QUALITY MEASUREMENT PLAN USING MONTE CARLO METHODS

A THESIS
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Faker ZOUAOUI

May, 1997

QUALITY MEASUREMENT PLAN USING MONTE CARLO METHODS

A THESIS

SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL

ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

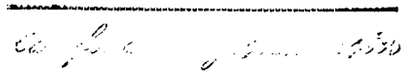
FOR THE DEGREE OF

MASTER OF SCIENCE

By

Faker Zouaoui

May, 1997



I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Ülkü Gürler(Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Asad Zaman(Co-supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Halim Doğrusöz

Approved for the Institute of Engineering and Sciences:

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Sciences

ABSTRACT

QUALITY MEASUREMENT PLAN USING MONTE CARLO METHODS

Faker Zouaoui
M.S. in Industrial Engineering
Supervisor: Assoc. Prof. Ülkü Gürler
Co-supervisor: Prof. Asad Zaman
May, 1997

This study considers the Quality Measurement Plan (QMP), a system implemented for reporting the quality assurance audit results to Bell system management. QMP is derived from a new Bayesian approach to the empirical Bayes problem for Poisson observations. It uses both the current and past data to compute estimates for the quality of the current production. The QMP estimator developed by Hoadley in 1981 is based on many complicated approximations. Sampling approaches such as the Gibbs sampler and Importance-sampling are alternative techniques that avoid these approximations and permit the computation of the quality estimates through Monte Carlo methods. Here we discuss the approaches and the algorithms for implementing some Monte Carlo-based approaches on the QMP model. We also show via simulation that although the QMP algorithm can be computationally more convenient, the sampling approaches mentioned above give more accurate estimates of current quality.

Key words: QMP, Hierarchical Bayes, Gibbs Sampler, Importance-Sampling, Substitution-Sampling.

ÖZET

MONTE-CARLO YÖNTEMLERINI KULLANARAK KALITE ÖLÇÜM PLANI

Faker Zouaoui
Endüstri Mühendisliği, Yüksek Lisans
Danışman: Doçent Ülkü Gürler
Ortak Danışman: Profesör Asad Zaman
Mayis, 1997

Bu çalışma, kalite güvencesi denetim sonuçlarının Bell sistem yönetimine aktarılması amacıyla yürütülen "Quality Measurement Plan (QMP)" (kalite ölçüm planı'nı) ele almaktadır. QMP, Poisson gözlemler için ampirik Bayes problemine yeni bir Bayes'ci yaklaşımdan türetilmiştir. Bu yöntem, ürün kalitesini tahmin etmek için hem şimdiki hem de geçmiş verileri kullanmaktadır. Hoadley'in 1981'de geliştirdiği QMP tahmin edicisi bir dizi karmaşık ve yaklaşık hesaplamalara dayanmaktadır. Gibbs örnekleyicisi ve önem örneklemesi gibi örnekleme yöntemleri bu yaklaşıklık gereğini ortadan kaldermakta ve kalite tahmin edicilerinin hesaplanmasında Monte-Carlo tekniklerinin kullanılmasına izin vermektedir. Bu çalışmada QMP modelinde Monte-Carlo tekniklerine dayalı bazı yaklaşımların denenmesi için algoritmalar ve yaklaşımlar tartışılmaktadır. Simülasyon sonuçlarımız göstermiştir ki, QMP algoritması hesaplama acısından daha etkin kullanılmış olmakla birlikte, örnekleme yaklaşımları mevcut kalitenin daha doğru bir tahminini vermektedir.

Anahtar Sözcükler: QMP, Hiyerarşik Bayes, Gibbs Örnekleyicisi, Önem Örneklemesi, Ikame Örneklemesi.

To my parents

AKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Asad Zaman due to his supervision, suggestions, and understanding to bring this thesis to an end.

I am especially indebted to Assoc. Prof. Ülkü Gürler for her invaluable guidance, and encouragement.

I would like to thank Prof. Halim Doğrusöz for showing keen interest to the subject matter and accepting to read and review this thesis.

I cannot fully express my gratitude and thanks to Çiğdem Gültekin and my parents for their morale support and encouragement.

I would also like to thank Souheyl Touhami, Alev Kaya, Maher Lahmar, Tijani Chahed, and Mehdi Bejar for their friendship and support.

Contents

1 Introduction **1**

2 The Quality Measurement Plan (QMP) **5**

2.1 Introduction 5

2.2 Statistical Foundations of QMP 6

2.2.1 The QMP Model 8

2.2.2 Posterior Distribution of Current Quality 10

2.3 Mathematical Derivation of QMP 13

2.3.1 Exact Solution 13

2.3.2 The QMP Formulae 14

2.4 Pros and Cons of the QMP 19

3 Monte Carlo-Based Approaches to Calculating Marginal densities **21**

3.1 Introduction 21

3.2 Monte Carlo Approaches 23

3.2.1 Substitution or Data-Augmentation Algorithm 23

3.2.2 Substitution Sampling 25

3.2.3 Gibbs Sampling 27

3.2.4 Relationship between Gibbs sampling and substitution sampling 29

3.2.5 The Rubin Importance Sampling Algorithm 31

3.3 Density Estimation 33

3.4 Sampling Issues 35

3.4.1 The Griddy-Gibbs Sampler 36

3.5 Convergence Issues 37

3.5.1 "Gelfand, Hills, Racine-Poon and Smith" methods 38

3.5.2 Tanner and Wong methods 39

3.5.3 The Gibbs Stopper 40

3.6 Gibbs Sampling with Improper Posteriors 41

3.7 Summary Discussion 43

4 QMP using Monte Carlo Methods 44

4.1 The Hierarchical Bayes Model 44

4.1.1 Model 1 44

4.1.2 Model 2 46

4.2 Monte Carlo Algorithms 48

4.2.1 Gibbs Sampler Algorithm 48

4.2.2 Importance-Sampling Algorithm 49

4.2.3 Implementation Issues 50

4.3 Simulation Study 53

4.3.1 Simulation Design 53

4.3.2 Simulation Results 54

5 Conclusion **56**

List of Figures

4.1 The mean of $\widehat{\theta}_6$ across iterations. 51

4.2 The posterior density of $\widehat{\theta}_6$. The dashed line represent iteration 5, and the solid line represent iteration 10. 52

4.3 The posterior density of $\widehat{\theta}_6$. The dashed line represent iteration 15-20, and the solid line represent iteration 10. 53

List of Tables

4.1 Equivalent Defects in Keys of Telephone sets - Shreveport 50

4.2 Simulation Results 55

Chapter 1

Introduction

Thousands of products are designed by Bell laboratories and produced by A. T. & T. (formerly called Western Electric). For quality control purposes, *audit*[1] samples are taken from each product and tested for defects. The main objective of the quality assurance department is to estimate the quality of the product from the sample, and to decide whether the product meets standard quality requirements or not.

The statistical foundations of the audit ingredients were developed by Shewhart, Dodge, and others, starting in the 1920's and continuing through to the middle 1950's. This work was documented in the literature in[41, 40, 10, 11] and evolved into the *T-rate* system. The basic idea behind the T-rate system is that observed quality results can be statistically compared to the standards, using a statistic called the T-rate.

For a given production period, let Q denote the total number of defects that are observed in all the inspections conducted on all the products. Because there are quality standards for each set of inspections on each product, it is possible to compute the standard mean and variance of Q under a given standard, denoted by $E(Q \mid S)$,and $V(Q \mid S)$. The T-rate is

[1] An audit is a highly structured system of inspections done on a sampling basis.

$$\text{T-rate} = \frac{E(Q \mid S) - Q}{\sqrt{V(Q \mid S)}}$$

It measures the difference between the observed result and its standard in units of statistical standard deviation. The T-rate is plotted in the control chart. The control limits of ± 2 are reasonable under the assumption that Q has an approximate normal distribution. Then the standard distribution of Q is the "standard normal", and excursions outside the control limits are rare under standard quality. For large audit samples, this approximation follows from the central limit theorem. As we shall see, the approximation is poor for small samples.

The advantage of the T-rate is its simplicity. It can be calculated manually. Exceptions can be identified by inspection. The fact that the T-rate has been used for so long is a testimonial to its advantages. However, the T-rate does have problems. The T-rate is not a direct measure of quality. A T-rate of -6 does not mean that quality is twice as bad as when the T-rate is -3. The T-rate is only a measure of statistical evidence with respect to the hypothesis of standard quality. To be specific, suppose the quality standard requires $\theta_t = 1$. If an estimate $\widehat{\theta}_t = 0.99$ has standard deviation 0.003, it is three standard deviations away from the null, and we will reject the null. However, the dereliction does not appear quantitatively serious. This is quite different from a situation where $\widehat{\theta}_t = 0.05$ and has standard deviation 0.05, and both are different from a case where $\widehat{\theta}_t = 0.5$ and the standard deviation is 0.25. In the first case we know for sure that the quality is slightly below normal. In the second we know for sure that the quality is way below normal, and in the third, we have a lot of uncertainty as to the real quality. In all three cases we reject the null hypothesis, and no further information is provided. Many other problems relating the T-rate system are described in detail in [22].

In the late seventies, research has been carried out to evaluate the application of modern statistical theories to quality assurance. An important idea is summarized in an article by Efron and Morris[13] which explains a paradox

discovered by Stein[43]. When you have samples from similar populations, the individual sample characteristics are not the best estimates of the individual population characteristics. Total error is reduced by shrinking the individual sample characteristics part way towards the grand mean over all samples. Efron and Morris used baseball batting averages to illustrate this point. But the problem of estimating percent defective in quality assurance is the same problem. And you are always concerned with similar populations-for example, the population of design-line telephones produced for each of several months. This idea was originally explored in Hoadley[21]. The idea has now evolved into the Quality Measurement Plan(QMP).

QMP was implemented throughout Western Electric in 1981 and by Bellcore in 1984. It is an Hierarchical Bayes(HB) approach to the control chart. It replaced the T-rate system described briefly above. Many of the advantages of the QMP relate to the disadvantages of the T-rate system. The main advantage is that unlike the T-rate, QMP uses past and current data to provide an inference about current quality not past quality. Hoadley[22] gives the rationale for changing the T-rate system.

Chapter 2 gives the statistical foundations of the QMP. The QMP model is described along with the QMP algorithm for estimating the posterior distribution of current quality. The methods developed by Hoadley in solving the QMP model might be the only methods available at that time in solving HB problems. Although they are computationally efficient, they are based on many complicated approximations and they may not work on all data sets. However, with the developments in computing power, many other techniques have emerged in the late 80's and 90's in order to calculate posterior distributions in HB problems. They are based on sampling approaches using Monte Carlo methods.

Chapter 3 gives a unified exposition of these techniques and evaluates their potential for HB problems. Most of the implementation issues such as the sampling and convergence issues are discussed thoroughly in this chapter.

Chapter 4, discusses how to implement the sampling approaches described

in chapter 3 on the QMP model. Some small modifications are added to the model in order to reformulate it in HB terms. Moreover, we show via simulation that the new implemented algorithms performed better than the existing QMP algorithm.

Finally in chapter 5, we summarize the different aspects of our study. We also underline future extensions of this work.

Chapter 2

The Quality Measurement Plan (QMP)

2.1 Introduction

This chapter is not intended to document QMP. We are only interested in the mathematics of the QMP and how it relates to the Hierarchical Bayes. Readers who are interested in the rationale for changing the rating system, the operating characteristics of QMP and its reporting format may refer to Hoadley[22, 23], and Bellcore[1]. Section 2.2 illustrates the statistical foundations of the QMP. Here we describe the QMP model and give the form of the posterior distribution of current quality. Section 2.3 shows that it is computationally impractical to derive the exact posterior distribution of current quality and gives a heuristic algorithm for QMP. Finally, Section 2.4 provides the pros and cons in developing the heuristic algorithm of the QMP.

2.2 Statistical Foundations of QMP

For the purpose of reporting quality of an audit results to management, the products are grouped into *rating classes.* The results of all the inspections associated with this rating class are aggregated over a time period called a *rating period.* In Bell laboratories a rating period is about six weeks and there are about eight rating periods per year. The defects assessed in each rating period are transformed into demerits or defectives or may remain as simple unweighted defects. In an audit based on demerits, each defect assessed is assigned a number of demerits: 100, 50, 10, or 1 for A, B, C, or D weight defects, respectively. In an audit based on defectives, all defects found in a unit of product are analyzed to determine if the unit is considered defective.

A complicating factor in the analyses of audit results is that defects, defectives, and demerits are different. But in fact they are not different; because, for statistical purposes, they can all be transformed into equivalent defects that have approximate Poisson distributions. Suppose we have a quality measure Q (Total defects, defectives, or demerits). Let E_s and V_s denote the standard mean (called expectancy) and variance of Q. So the T-rate is is T= $(E_s - Q)/\sqrt{V_s}$.

Now define

$$X = \text{equivalent defects} = \frac{Q}{V_s/E_s},$$

and

$$\begin{aligned} e &= \text{equivalent expectancy} \\ &= \text{standard mean of } X. \\ &= \frac{E_s}{V_s/E_s} \\ &= \frac{E_s^2}{V_s}. \end{aligned}$$

If all defects have Poisson distributions and are occurring at θ times the standard rate, then it can be shown that

$$E[X \mid \theta] = V[X \mid \theta] = e\theta.$$

Hence, X has an approximate Poisson distribution with mean $e\theta$.

As an example, consider the demerits case. The total number of demerits has the general form

$$D = \sum w_i X_i,$$

where the w_i's are known weights and the X_i's have Poisson distributions. Assume that the mean of X_i is $e_i\theta$, where e_i is the standard mean of X_i and θ is the population quality expressed on an index scale. So $\theta = 2$ means that all types of defects are occuring at twice the rate expected.

The mean and varinace of D are

$$\begin{aligned} E(D) &= \sum w_i E(X_i) \\ &= \sum w_i (e_i\theta) \\ &= \theta E_s, \end{aligned}$$

and

$$\begin{aligned} V(D) &= \sum w_i^2 V(X_i) \\ &= \sum w_i^2 (e_i\theta) \\ &= \theta V_s, \end{aligned}$$

where E_s and V_s are the standard mean and variance, respectively, of D. These are the numbers that would be published in the official list of standards called the Master Reference list.

The mean and variance of equivalent defects, X, are

$$\begin{aligned} E(X) &= \frac{E(D)}{V_s/E_s} \\ &= \frac{\theta E_s}{V_s/E_s} \\ &= \theta e. \end{aligned}$$

and

$$\begin{aligned} V(X) &= \frac{V(D)}{[V_s/E_s]^2} \\ &= \frac{\theta V_s E_s}{V_s^2} \\ &= \theta e. \end{aligned}$$

The mean and variance of X are equal; so, X has an approximate Poisson distribution with mean $e\theta$. Of course, it is not exact, because X is not always integer valued. A similar analysis works for the defectives case. So, any aggregate of demerits, defectives, or defects can be transformed into equivalent defects. Just use the standard expectancy and variance as illustrated above for demerits.

2.2.1 The QMP Model

For rating period t, let x_t = equivalent defects in the audit sample, e_t = equivalent expectancy of the audit sample, and θ_t = population index as defined previously. Based on our previous discussion , we assume that

$$x_t \mid \theta_t \sim \text{Poisson}\,(e_t\theta_t)\,.$$

For reasons that are partly statistical and partly administrative, Hoadley

decided to restrict his use of past data to five periods. The main administrative reason is that the T-rate system used the past five periods.

A consequence of using only six periods of data is that no useful inference can be made about the possible complex structure in the stochastic process of θ_t's. So we assume simply that the θ_t's are a random sample from an unknown distribution called the process distribution. Furthermore, six observations are not enough to make fine inferences about the family of this unknown distribution. So for mathematical simplicity we assume it to be a gamma distribution with unknown mean θ (called the process average) and variance γ^2(called the process variance). The gamma distribution is used because it is the natural conjugate prior to the Poisson distribution and it is a reasonable parametric model of a unimodal distribution on the nonnegative real numbers. The choice of a unimodal distribution reflects the experience that usually many independent factors affect quality; so there is a central limit theorem effect.

The model so far is an empirical Bayes model. The parameter of interest is the current population index, θ_T, which has a distribution called the process distribution. Bayesians would call it the prior distribution if it were known. But we must use all the data to make an inference about the unknown process distribution. So, the model is called empirical Bayes.

Efron and Morris[12] take a classical approach to the Empirical Bayes model. They use classical methods of inference for the unknown process distribution. QMP is based on a Bayesian approach to the empirical Bayes model. Hence, the model is called Hierarchical Bayes (HB). Each product has its own process mean and variance. These vary from product to product. By analyzing many products, we can model this variation by a prior distribution for (θ, γ^2).

Summarizing, the QMP model is

$$\begin{aligned} x_t \quad &| \quad \theta_t \sim \text{Poisson}\left(e_t\theta_t\right), t = 1, \ldots, T, \\ \theta_t \quad &\sim \quad \text{Gamma}\left(\frac{\theta^2}{\gamma^2}, \frac{\gamma^2}{\theta}\right), \end{aligned}$$

$$\left(\theta, \gamma^2\right) \sim \text{prior distribution } \rho\left(\theta, \gamma^2\right).$$

This is a full Bayesian model. It specifies the joint distribution of all variables. The quality rating in QMP is based on the posterior distribution of θ_T given $\mathbf{x} = (x_1, \ldots, x_T)$.

2.2.2 Posterior Distribution of Current Quality

We show in the next section that it is computationally impractical to derive the exact posterior distribution of θ_T. The best we can do is approximate the posterior mean and variance of θ_T.

The posterior mean and variance are derived in Section IV of Hoadley[22]. A brief discussion of how to derive them is given in the next section. The posterior mean is

$$\begin{aligned} \widehat{\theta}_T &= E\left(\theta_T \mid \mathbf{x}\right) \\ &= \widehat{\omega}_T \widehat{\theta} + \left(1 - \widehat{\omega}_T\right) I_T. \end{aligned}$$

where

$$\begin{aligned} I_T &= x_T / e_T, \\ \widehat{\theta} &= E\left(\theta \mid \mathbf{x}\right), \\ \widehat{\omega}_T &= E\left(\omega_T \mid \mathbf{x}\right), \\ \omega_T &= \frac{\theta / e_T}{\theta / e_T + \gamma^2}. \end{aligned}$$

The posterior mean, $\widehat{\theta}_T$, is a weighted average between the estimated process average, $\widehat{\theta}$, and the defect index, I_T, of the current sample. It is the dynamics of the weight, $\widehat{\omega}_T$, that makes the Bayes estimate work well. For any t, the sampling variance of I_T is

$$\begin{aligned} V(I_t \mid \theta_t) &= V\left(\frac{x_t}{e_t} \mid \theta_t\right) \\ &= \frac{1}{e_t^2}(e_t\theta_t) \\ &= \theta_t/e_t. \end{aligned}$$

The expected value of this is

$$E[\theta_t/e_t] = \theta/e_t.$$

So the weight, ω_T, is

$$\frac{\text{[expected sampling variance]}}{\text{[expected sampling variance]} + \text{[process variance]}},$$

If the process is relatively stable, then the process variance is relatively small and the weight is mostly on the process average; but if the process is relatively unstable, then the process variance is relatively large and the weight is mostly on the current sample index. The reverse is true of the sampling variance. In other words, ω_T, is a monotonic function of the ratio of expected sampling variance to process variance. The posterior variance of θ_T is

$$V_T = (1-\widehat{\omega}_T)\,\widehat{\theta}_T/e_T + \widehat{\omega}_T^2 V(\theta \mid \mathbf{x}) + \left(\widehat{\theta} - I_T\right)^2 V(\omega_T \mid \mathbf{x}).$$

If the process average and variance were known, then the posterior variance of θ_T would be $(1-\omega_T)\,\widehat{\theta}_T/e_T$. So the first term is just an estimate of this. But since the process average and variance are not known, the posterior variance has two additional terms. One contains the posterior variance of the process average and the other contains the posterior variance of the weight.

If the process average and variance were known, then the posterior distribution would be gamma (see the next section). So we approximate the posterior

distribution with a gamma fitted by the method of moments. The parameters of the fitted gamma are

$$\begin{aligned} \upsilon &= \text{Shape parameter} = \frac{\widehat{\theta}_T^2}{V_T}, \\ \tau &= \text{Scale parameter} = \frac{V_T}{\widehat{\theta}_T}, \end{aligned}$$

and the posterior cumulative distribution is

$$\Pr\left(\theta_T \leq y \mid \mathbf{x}\right) = G_\upsilon(y/\tau)$$

The QMP formulae for the above terms are given in Section 2.3.2. These are derived by Hoadley[22, 23]. The actual scheme used by Hoadley begins with the formulae above, but with substantial developments of several types. He used some moments of the marginal distribution of x_T which is a negative binomial distribution (See the next section) to calculate weighted average estimates of the hyperparameters. Moreover, he calculated some empirical prior distributions for the hyperparameters. He did not use the exact forms of these distributions, but just some of their characteristics such as mean, variance, and mode. A thorough discussion on these developments is given in Hoadley[22].

The main objective of Hoadley in developing the QMP formulae at that time is to sell them to the management and to the engineers. Many of the developments in his estimation procedures were in fact ad-hoc, these are listed in Section 2.4. These were quite complex, but were a product of the necessity to be sufficiently better than the earlier quality management plan in effect as to be bureaucratically acceptable, and to be easy to compute.

2.3 Mathematical Derivation of QMP

2.3.1 Exact Solution

We are interested in the posterior distribution of θ_T given $\mathbf{x}$, for the model described above. For mathematical convenience let us define the hyperparameters α and β as: $\alpha = \frac{\theta^2}{\gamma^2}$ and $\beta = \frac{\gamma^2}{\theta}$.

Given $x_t \mid \theta_t \sim \text{Poisson}(e_t\theta_t)$ and $\theta_t \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, then the posterior density of θ_t is also gamma, and the marginal density of x_t is negative binomial. The following calculation factorizes the joint density of x_t and θ_t into the product of the marginal of x_t and the posterior of θ_t given x_t.

$$\begin{aligned}
f(x_t, \theta_t) &= f(x_t \mid \theta_t) \times f(\theta_t) \\
&= \{P(\theta_t e_t) \times G(\alpha, \beta)\} \\
&= \left(\frac{(\theta_t e_t)^{x_t} e^{-\theta_t e_t}}{x_t!}\right) \times \left(\frac{\theta_t^{\alpha-1} e^{-\theta_t/\beta}}{\beta^\alpha \Gamma(\alpha)}\right) \\
&= \left(\frac{e_t^{x_t} (1/\beta)^\alpha \Gamma(x_t + \alpha)}{x_t! \Gamma(\alpha) (1/\beta + e_t)^{x_t+\alpha}}\right) \times \left(\frac{\theta_t^{x_t+\alpha-1} e^{-(1/\beta+e_t)\theta_t}}{(1/\beta + e_t)^{-(x_t+\alpha)} \Gamma(x_t + \alpha)}\right) \\
&= f(x_t) \times f(\theta_t \mid x_t) \\
&= \left\{NB(\alpha, \beta) \times G\left(x_t + \alpha, (e_t + 1/\beta)^{-1}\right)\right\}
\end{aligned}$$

Now,

$$\Pr(\theta_T \leq y \mid \mathbf{x}) = \int_0^\infty \int_0^\infty \Pr(\theta_T \leq y \mid \alpha, \beta, x_T) \rho(\alpha, \beta \mid \mathbf{x}) \, d\alpha \, d\beta,$$

where $\rho(\alpha, \beta \mid \mathbf{x})$ is the posterior distribution of α, β given $\mathbf{x}$.

We know from the above calculation that the distribution of θ_T given α, β and x_T is gamma; so $\Pr(\theta_T \leq y \mid \alpha, \beta, x_T)$ can be expressed in terms of an incomplete gamma function.

By Bayes theorem,

$$\rho(\alpha,\beta \mid \mathbf{x}) = \frac{\rho(\alpha,\beta \mid \mathbf{x})\, L(\mathbf{x} \mid \alpha,\beta)}{\int_0^\infty \int_0^\infty \rho(\alpha,\beta \mid \mathbf{x})\, L(\mathbf{x} \mid \alpha,\beta)\; d\alpha\, d\beta},$$

where $\rho(\alpha,\beta \mid \mathbf{x})$ is the prior density of (α,β) and $L(\mathbf{x} \mid \alpha,\beta)$ is the likelihood function. We know from above that x_t given α,β is negative binomial. Hence,

$$L(\mathbf{x} \mid \alpha,\beta) = \prod_{t=1}^{T} \frac{e_t^{x_t}\,(1/\beta)^\alpha\,\Gamma(x_t+\alpha)}{x_t!\Gamma(\alpha)\,(1/\beta+e_t)^{x_t+\alpha}}$$

So the posterior distribution of θ_T is a complex triple integral that has to be inverted to compute the QMP box chart. The posterior mean and variance of θ_T can be expressed in terms of several double integrals. There are more than 1,000 rating classes that have to be analyzed each period, so computational efficiency is important. This is why heuristic algorithms were developed for the QMP model.

2.3.2 The QMP Formulae

Two heuristic algorithms have been implemented for QMP. The first was derived by Hoadley[22] and the second, which is currently used at Bellcore, was derived by Hoadley[23]. For completeness, we state the formulae from Hoadley[23] that are needed to implement the QMP estimator. The formulae look complex, but they are algebraic and easily programmable.

As explained before, the QMP model has a prior distribution on (θ,γ^2). The parameters of this prior that we use are θ_0 = prior mean of θ, ν_0 = prior variance of θ, γ_0^2 = prior mean of γ^2, and $\gamma_{\max}^2$ is defined by $\Pr\{\gamma^2 \le \gamma_{\max}^2\} = .95$. We require a technical constraint $\nu_0 > \gamma_0^2$, which is caused by some of the prior information being implemented in the algorithm as artificial data:

$$e_0 \;=\; \text{prior expectancy}$$

$$= \theta_0 / \left(\nu_0 - \gamma_0^2\right)$$

and

$$\begin{aligned} x_0 &= \text{prior defects} \\ &= \theta_0 e_0. \end{aligned}$$

The default priors used in the current Bellcore implementation of QMP are $\theta_0 = 1$, $\nu_0 = 3.05$, $\gamma_0^2 = .55$, and $\gamma_{\max}^2 = 2.2$. These were arrived at by analysis of factory quality-audit data for many products.

The audit data for $t = 1, 2, \ldots, T$ is the following:

$$\begin{aligned} Q_t &= \text{Attribute quality measure in the sample period } t \\ &\quad \text{(total defects, defectives, or demerits),} \\ E_{st} &= \text{Expected value of } Q_t \text{ given standard quality,} \\ V_{st} &= \text{Sampling variance of } Q_t \text{ given standard quality.} \end{aligned}$$

For each period compute the following:

Equivalent defects:

$$x_t = \frac{Q_t}{V_{st}/E_{st}}$$

Equivalent expectancy:

$$e_t = \frac{E_{st}^2}{V_{st}}.$$

For $t = 0, \ldots, T$, compute the following:

Sample index:

$$I_t = x_t/e_t,$$

Weighting factors for computing process average and variance:

$$\begin{aligned} b_t &= \gamma_0^2 + \theta_0/e_t, \\ \phi_0 &= 2 + 6\gamma_0^2/\theta_0^2 \\ f_t &= 1/b_t, \\ g_t &= 1/\left(\phi_0 b_t^2 - 4\gamma_0^2 b_t/\left[\theta_0 e_t\right]\right). \end{aligned}$$

Corresponding weights:

$$\begin{aligned} p_t &= f_t/\sum_{t=0}^{T} f_t, \\ q_t &= g_t/\sum_{t=0}^{T} g_t. \end{aligned}$$

Over all periods compute the following:

Process average:

$$\widehat{\theta} = \textstyle\sum_{t=0}^{T} p_t I_t,$$

Average sampling variance:

$$\sigma^2 = \textstyle\sum_{t=1}^{T} q_t \left(\widehat{\theta}/e_t\right),$$

Total observed variance:

$$V = \textstyle\sum_{t=1}^{T} q_t \left(I_t - \widehat{\theta}\right)^2,$$

Shape parameter for the sampling distribution of V:

$$a_1 = \frac{\left[\sum_{t=1}^{T} q_t \left(\theta_0/e_t\right)\right]^2}{\sum_{t=1}^{T} q_t^2 \left[2\left(\theta_0/e_t\right)^2 + \theta_0/e_t^3\right]},$$

Shape parameter for the prior distribution of $\omega = \sigma^2/\left(\sigma^2 + \gamma^2\right)$:

$$a_0 = \frac{\ln\left(20\right)}{\ln\left[1 + \gamma_{\max}^2/\sigma^2\right]},$$

Shape parameter of the posterior distribution of ω:

$$a = a_0 + a_1,$$

Adjusted total variance:

$$S^2 = \left(a_1/a\right) V,$$

Adjusted variance ratio:

$$R = S^2/\sigma^2,$$

Bayes adjustment factor used to keep the estimator process variance positive:

$$\begin{aligned}
B &= \textstyle\sum_{i=1}^{m} T\left(i\right),\ T\left(0\right) = 1, \\
T\left(i\right) &= T\left(i-1\right)\left[\frac{aR}{a+i}\right], \\
&\quad T\left(m\right) \text{ is the term in which either} \\
T\left(m\right) &> 10^7 \text{ or } T\left(m\right) < 10^{-7}, \\
F &= 1 + \left(1/B\right), \\
\text{If } R &= 0, F \text{ is not needed.}
\end{aligned}$$

Posterior variance of ω:

$$G = \begin{cases} \frac{\left[\frac{(a+1)}{aR} - F + 1 - \frac{1}{FR}\right]}{FR} & \text{if} \quad R > 0, \\ \frac{a}{(a+2)(a+1)^2} & \text{if} \quad R = 0, \end{cases}$$

Current sampling variance:

$$\sigma_T^2 = \hat{\theta}/e_T,$$

Sampling variance ratio:

$$r_T = \sigma_T^2/\sigma^2,$$

Process variance:

$$\hat{\gamma}^2 = \begin{cases} FS^2 - \sigma^2 & \text{if} \quad R > 0, \\ \sigma^2/a & \text{if} \quad R = 0, \end{cases}$$

Average shrinkage weight:

$$\hat{\omega} = \sigma^2/\left(\sigma^2 + \hat{\gamma}^2\right),$$

Shrinkage weight for the current period:

$$\hat{\omega}_T = \sigma_T^2/\left(\sigma_T^2 + \hat{\gamma}^2\right),$$

Best measure of current quality:

$$E\left(\theta_T \mid \mathbf{x}\right) = \hat{\omega}_T\hat{\theta} + \left(1 - \hat{\omega}_T\right) I_T = \hat{\theta}_T,$$

Posterior variance of current quality:

$$\begin{aligned} V(\theta_T \mid \mathbf{x}) &= (1-\widehat{\omega}_T)\,\widehat{\theta}_T/e_T + \widehat{\omega}_T^2 \sum_{t=0}^{T} p_t^2 \left[\widehat{\gamma}^2 + \frac{\widehat{\theta}}{e_t}\right] \\ &\quad + \frac{r_T^2\left(\widehat{\theta} - I_T\right)^2}{\left[(r_T-1)\,\widehat{\omega}+1\right]^4} G \\ &= V_T. \end{aligned}$$

2.4 Pros and Cons of the QMP

Many of the advantages of QMP relate to the disadvantages of the T-rate system. QMP does not use runs criteria, but uses the actual equivalent defects observed. It uses past and current data to make an inference about current quality not past quality. For example, under QMP a rating class is Below Normal if the posterior probability that the product is substandard[1] exceeds 0.99. However, under the T-rate system a rating class is Below Normal if the current T-rate is below -3 and at least one of the previous five T-rates is below -2. The T-rate system uses numerous patchwork rules to exploit information in the past history of the production process such as the rule of Below Normal. But the information is not used in an efficient way. A situation where there is long past experience on the same or similar parameters is ideally suited to the use of Bayesian methods. Another advantage of the QMP on the T-rate system is that it provides a lower producer's risk and consequently a more accurate list of exceptions (see Hoadley[22]). Finally, unlike the T-rate statistic, the QMP report format met the needs of the operating companies.

From the previous section it should be clear that the Hierarchical Bayes (HB) estimator is quite difficult to obtain. For this reason, applications of HB estimation have been very few. See the study by Nebebe and Stroud[29]. Approximating the posterior resulting from the HB analysis requires a lot of numerical integration. There are a number of different ad-hoc methods for carrying out such approximations, each adapted to particular applications.

[1]Substandard means $\theta_T > 1$.

Hoadley used a technique like this to compute the QMP estimator. In the past this was the only method of obtaining approximations to HB estimators. However, the invention of Monte Carlo approaches such as Gibbs and substitution sampling for computing HB estimators has rendered this type of approximation considerably less useful than before. So it is probably preferable to calculate the HB estimates by the Monte Carlo techniques rather than approximate it in some ad-hoc fashion which is typically quite complex in itself like in our case.

In the next section, we will describe three Monte Carlo-based approaches put forward in the literature for calculating marginal densities (or posterior densities in HB applications) before applying these techniques to the QMP model.

Chapter 3

Monte Carlo-Based Approaches to Calculating Marginal densities

3.1 Introduction

Metropolis, Rosenbluth, Rosenbluth, Teller and Teller[28] introduced a Monte Carlo-type algorithm to investigate the equilibrium properties of large systems of particles, such as molecules in a Gas. Hastings[20] used the Metropolis algorithm to sample form certain distributions; for example Poisson, standard normal, and random orthogonal matrices. Besag[2] studied the associated Markov field structure. Geman and Geman[17] illustrated the use of a version of this algorithm that they called the Gibbs sampler in the context of image reconstruction. More recently, Tanner and Wong[44] developed a framework in which Gibbs sampler algorithms can be used to calculate posterior distributions; and Carlin, Gelfand, and Smith[4]; Carlin and Poison[5]; Gelfand, Hills, Racine-Poon, and Smith[14]; Gelfand and Smith[15]; Gelfand, Smith, and Lee[16]; Li[27]; Verdinelli and Wasserman[45]; and Zeger and Karim[49] used the Gibbs sampler to perform Bayesian Computations in various important statistical

problems.

In Section 3.2, we discuss three alternative approaches put forward in the literature for calculating marginal densities via Monte Carlo (or sampling) algorithms. These are (variants of) the data-augmentation algorithm described by Tanner and Wong[44], the Gibbs sampler algorithm introduced by Geman and Geman[17], and the form of importance-sampling algorithm by Rubin[38, 37]. We note that the Gibbs sampler has been widely taken up in the image-processing literature and in other large-scale models such as neural networks and expert systems, but in general potential for more conventional statistical problems seem to have been overlooked. As we show, there is a close relationship between the Gibbs sampler and the substitution or data augmentation algorithm. We generalize the latter and show that it is as efficient as the Gibbs sampler, and potentially more efficient, given the availability of distinct conditional distributions. We note that a consequence of the relationship between the two algorithms, the convergence results established by Geman and Geman[17] are applicable to the generalized substitution algorithm. Both the substitution and the Gibbs sampler algorithms are iterative Monte Carlo procedures. However, we see that an importance sampling algorithm based on that of Rubin[38, 37] provides noniterative Monte Carlo integration approach to calculating marginal densities.

In Section 3.3, we consider the problem of calculating a final form of marginal density from the final sample produced by either the substitution or Gibbs sampling algorithms. In Section 3.4, we propose several methods for sampling from *nonconjugate* conditional distributions. Moreover, some references are given for the efficient generation of variates from some known distributions. In Section 3.5, we propose some techniques useful for assessing convergence of the substitution or Gibbs sampling algorithms. In Section 3.6, we illustrate the effect of improper priors on Gibbs sampling. Finally, in Section 7 we provide a summary discussion.

3.2 Monte Carlo Approaches

In the sequel, we assume that we are dealing with real, possibly vector-valued random variables having a joint distribution whose density function is strictly positive over the (product) sample space. This ensures that knowledge of all full conditional distributions uniquely defines the joint density (e.g., see Besag[2]). Throughout,We assume the existence of densities with respect to either Lebesgue or counting measure, as appropriate, for all marginal and conditional distributions. The terms distribution and density are therefore used interchangeably. Densities are denoted by brackets, so joint, conditional, and marginal forms, for example appear as $[X, Y]$, $[X \mid Y]$, and $[X]$. Multiplication is denoted by $*$; for example, $[X, Y] = [X \mid Y] * [Y]$. The process of integration is denoted by forms such as $[X \mid Y] = \int [X \mid Y, Z, W] * [Z \mid W, Y] * [W \mid Y]$, with the convention that all variables appearing in the integrand but not in the resulting density have been integrated out.

3.2.1 Substitution or Data-Augmentation Algorithm

The substitution algorithm for finding fixed-point solutions to certain classes of integral equations is a standard mathematical tool that has received considerable attention in the literature (e.g., see Rall[31]). Its potential utility in statistical problems of the kind we are concerned with was observed by Tanner and Wong[44] (who called it the data-augmentation algorithm). Briefly reviewing the essence of their development using the notation introduced previously, we have

$$[X] = \int [X \mid Y] * [Y] \tag{3.1}$$

and

$$[Y] = \int [Y \mid X] * [X] \tag{3.2}$$

so substituting (3.2) into (3.1) gives

$$\begin{aligned} [X] &= \int [X \mid Y] * \int [Y \mid X'] [X'] \\ &= \int h(X, X') * [X'] \end{aligned} \tag{3.3}$$

Where $h(X, X') = \int [X \mid Y] * [Y \mid X']$, with X' appearing as a dummy argument in (3.3), and of course $[X] \equiv [X']$.

Now suppose that on the right side of (3.3), $[X']$ were replaced by $[X]_i$, to be thought as an estimate of $[X] \equiv [X']$ arising at the ith stage of an iterative process. Then, (3.3) implies that for some $[X]_{i+1}$, $[X]_{i+1} = \int h(X, X') * [X']_i = I_h[X]_i$, in a notation making explicit that I_h is the integral operator associated with h. Exploring standard theory of such integral operators, Tanner and Wong[44] showed that under mild regularity conditions this iterative process has the following properties (with obviously analogous results for $[Y]$).

TW1(uniqueness). The true marginal density, $[X]$, is the unique solution to (3.3).

TW2 (convergence). For almost any $[X]_0$, the sequence $[X]_1, [X]_2, \ldots$ defined by $[X]_{i+1} = I_h[X]_i (i = 0, 1, \ldots)$ converges monotonically in L_1 to $[X]$.

TW (rate). $\int |[X]_i - [X]| \to 0$ geometrically in i.

Extending the substitution algorithm to three random variables X, Y, and Z, we may write [analogous to (3.1) and (3.2)]

$$[X] = \int [X, Z \mid Y] * [Y] \tag{3.4}$$

$$[Y] = \int [Y, X \mid Z] * [Z] \tag{3.5}$$

$$[Z] = \int [Z, Y \mid X] * [X] \tag{3.6}$$

Substitution of (3.6) into (3.5) and then (3.5) into (3.4) produces a fixed-point equation analogous to (3.3). A new h function arises with associated integral operator I_h, and hence TW1, TW2, and TW3 continue to hold in this extended setting. Extension to k variables is straightforward.

3.2.2 Substitution Sampling

Returning to (3.1) and (3.2), suppose that $[X \mid Y]$ and $[Y \mid X]$ are available in the sense defined at the beginning of the previous section. For an arbitrary (possibly degenerate) initial density $[X]_0$ draw a single $X^{(0)}$ from $[X]_0$. Given $X^{(0)}$, since $[Y \mid X]$ is available draw $Y^{(1)} \sim \left[Y \mid X^{(0)}\right]$, and hence from (3.2) the marginal distribution of $Y^{(1)}$is $[Y]_1 = \int [Y \mid X] * [X]_0$. Now, complete a cycle by drawing $X^{(1)} \sim \left[X \mid Y^{(1)}\right]$. Using (3.1), we then have $X^{(1)} \sim [X]_1 = \int [X \mid Y] * [Y]_1 = \int h(X, X') * [X']_0 = I_h [X]_0$. Repetition of this cycle produces $Y^{(2)}$ and $X^{(2)}$, and eventually, after i iterations, the pair $\left(X^{(i)}, Y^{(i)}\right)$ such that $X^{(i)} \xrightarrow{d} X \sim [X]$, and $Y^{(i)} \xrightarrow{d} Y \sim [Y]$, by virtue of TW2. Repetition of this sequence m times each to the ith iteration generates m iid pairs $(X_j^{(i)}, Y_j^{(i)})$ $(j = 1, \ldots, m)$. We call this generation scheme *substitution sampling*. Note that though we have independence across j, we have dependence within a given j.

If we terminate all repetitions at the ith iteration, the proposed density estimate of $[X]$ (with an analogous expression for $[Y]$) is the Monte Carlo integration

$$\left[\widehat{X}\right]_i = \frac{1}{m} \sum_{j=1}^{m} \left[X \mid Y_j^{(i)}\right]. \tag{3.7}$$

Note that the $X_j^{(i)}$ are not used in (3.7).

We note that this version of the substitution-sampling algorithm differs slightly from the imputation-posterior algorithm of Tanner and Wong[44]. At each iteration $l(l = 1, 2, \ldots, i)$,they proposed creation of the mixture density

estimate , $[\hat{X}]_l$, of the form in (3.7), with subsequent sampling from $[\hat{X}]_l$ to begin the next iteration. This mechanism introduces the additional randomness of equally likely selection from the $Y_j^{(l)}$ before obtaining an $X^{(l)}$. We suspect this sampling with replacement of the $Y^{(l)}$ was introduced to allow m to vary across iterations, which may reduce computational effort.

The L_1 convergence of $[\widehat{X}]_i$ to $[X]$ is most easily studied by writing

$$\int \left|\left[\widehat{X}\right]_i - [X]\right| \leq \int \left|\left[\widehat{X}\right]_i - [X]_i\right| + \int \left|[X]_i - [X]\right|.$$

The second term on the right side can be made arbitrarily small as $i \to \infty$, as a consequence of TW2. The first term on the right can be made arbitrarily small as $m \to \infty$, since $[\widehat{X}]_i \xrightarrow{p} [X]_i$ for almost all X (Glick[19]).

Extension of the substitution-sampling algorithm to more than two random variables is straightforward. We illustrate using the three-variable case, assuming the three conditional distributions in (3.4)-(3.6) are available. Taking an arbitrary starting marginal density for X, say $[X]_0$, we draw $X^{(0)} \sim [X]_0$, $\left(Z^{(0)'}, Y^{(0)'}\right) \sim \left[Z, Y \mid X^{(0)}\right]$, $\left(Y^{(1)}, X^{(0)'}\right) \sim \left[Y, X \mid Z^{(0)'}\right]$, and finally $\left(X^{(1)}, Z^{(1)}\right) \sim \left[X, Z \mid Y^{(1)}\right]$. A full cycle of the algorithm (i.e., to generate $X^{(1)}$ starting from $X^{(0)}$) thus requires six generated variates, rather then the two we saw earlier. Repeating such a cycle i times produces $\left(X^{(i)}, Y^{(i)}, Z^{(i)}\right)$. The aforementioned theory ensures that $X^{(i)} \xrightarrow{d} X \sim [X]$, $Y^{(i)} \xrightarrow{d} Y \sim [Y]$, and $Z^{(i)} \xrightarrow{d} Z \sim [Z]$. If we repeat the entire process m times we obtain iid $\left(X_j^{(i)}, Y_j^{(i)}, Z_j^{(i)}\right)$ $(j = 1, ..., m)$ (independent between, but not within, $j's$). Note that implementation of the substitution-sampling algorithm does not require specification of the full joint distribution. Rather, what is needed is the availability of $[X, Z \mid Y], [Y, X \mid Z]$, and $[Z, Y \mid X]$. Of course, in many cases sampling from, say, $[X, Z \mid Y]$ requires, for example, $[X|Y, Z]$ and $[Y \mid X]$, that is, the availability of a full conditional and a reduced conditional distribution. Paralleling (3.7), the density estimator of $[X]$ becomes

$$\left[\widehat{X}\right]_i = \frac{1}{m}\sum_{j=1}^{m}\left[X \mid Y_j^{(i)}, Z_j^{(i)}\right], \tag{3.8}$$

With analogous expressions for estimating $[Y]$ and $[Z]$. L_1 convergence of 3.8 to $[X]$ again follows.

For k variables, $U_1, ..., U_k$, the substitution-sampling algorithm requires $k(k-1)$ random variate generations to complete a cycle. If we run m sequences out to the ith iteration [$mik(k-1)$ random generations] we obtain m iid k tuples $\left(U_{1,j}^{(i)}, ..., U_{k,j}^{(i)}\right)$ $(j = 1, ..., m)$, with the density estimator for $[U_s]$ $(s = 1, .., k)$ being

$$\left[\hat{U_s}\right]_i = \frac{1}{m}\sum_{j=1}^{m}\left[U_s \mid U_t = U_{t,j}^{(i)}; t \neq s\right] \tag{3.9}$$

3.2.3 Gibbs Sampling

Suppose that we write (3.4)-(3.6) in the form

$$\begin{aligned}
[X] &= \int [X \mid Z, Y] * [Z \mid Y] * [Y] \\
[Y] &= \int [Y \mid X, Z] * [X \mid Z] * [Z] \\
[Z] &= \int [Z \mid Y, X] * [Y \mid X] * [X]
\end{aligned} \tag{3.10}$$

Implementation of substitution sampling requires the availability of all six conditional distributions on the right side of (3.10), rarely the case in our applications. As noted at the beginning of Section 3.2, the full conditional distributions alone, $[X \mid Z, Y]$, $[Y \mid X, Z]$, and $[Z \mid Y, X]$, uniquely determine the joint distribution (and hence the marginal distributions) in the situations under study. An algorithm for extracting the marginal distributions from these full conditional distributions was formally introduced by Geman and Geman[17] and is known as Gibbs sampler. An earlier article by Hastings[20] developed

essentially the same idea and suggested its potential for numerical problems arising in statistics.

The Gibbs sampler was developed and has mainly being applied in the context of complex stochastic models involving very large numbers of variables, such as image reconstruction, neutral networks, and expert systems. In these cases, direct specification of a joint distribution is typically not feasible. Instead, the set of full conditionals is specified, usually by assuming that an individual full conditional distribution only depends on some "neighborhood" subset of the variables. More precisely, for the set of variables $U_1, U_2, ..., U_k$,

$$[U_i \mid U_j; j \neq i] \equiv [U_i \mid U_j; j \in S_i], \ i = 1, ..., k, \tag{3.11}$$

where S_i is a small neighborhood subset of $\{1, 2, ..., k\}$. A crucial question is under what circumstances the specification (3.11) uniquely determines the joint distribution. The answer is taken up in great detail by Geman and Geman[17], involving concepts such as graphs, neighborhood systems, cliques, Markov random fields, and Gibbs distributions. Section 3.5 provides also some references to answer this crucial question.

Gibbs sampling is a Markovian updating scheme that proceeds as follows. Given an arbitrary starting set of values $U_1^{(0)}, U_2^{(0)}, \ldots, U_k^{(0)}$, we draw $U_1^{(1)} \sim \left[U_1 \mid U_2^{(0)}, \ldots, U_k^{(0)}\right]$, $U_2^{(1)} \sim \left[U_2 \mid U_1^{(1)}, U_3^{(0)}, \ldots, U_k^{(0)}\right]$, and so on, up to $U_k^{(1)} \sim \left[U_k \mid U_1^{(1)}, \ldots, U_{k-1}^{(1)}\right]$. Thus each variable is visited in the natural order and a cycle in this scheme requires K random variate generations. After i such iterations we could arrive at $\left(U_1^{(i)}, \ldots, U_k^{(i)}\right)$. Under mild conditions, Geman and Geman showed that the following results hold.

GG1 (convergence). $\left(U_1^{(i)}, \ldots, U_k^{(i)}\right) \xrightarrow{d} [U_1, \ldots, U_k]$ and hence for each s, $U_s^{(i)} \xrightarrow{d} U_s \sim [U_s]$ as $i \to \infty$.

In fact, a slightly stronger result is proven. Rather than requiring that each variable be visited in repetitions of the natural order, convergence still follows under any visiting scheme, provided that each variable is visited infinitely often

(io).

GG2 (rate). Using the sup norm, rather than L_1 norm, the joint density of $\left(U_1^{(i)}, \ldots, U_k^{(i)}\right)$ converges to the true joint density at a geometric rate in i, under visiting in the natural order. A minor adjustment to the rate is required for an arbitrary io visiting scheme.

GGS (ergodic theorem). For any measurable function T of $U_1, \ldots, U_k$ whose expectation exists, $\lim_{i\to\infty} \frac{1}{i} \sum_{l=1}^{i} T\left(U_1^{(l)}, \ldots, U_k^{(l)}\right) \stackrel{a.s.}{\to} E\left(T\left(U_1, \ldots, U_k\right)\right)$.

As in the previous section, Gibbs sampling through m replications of the aforementioned i iterations (mik random variate generations) produces m iid k tuples $\left(U_{1j}^{(i)}, ..., U_{kj}^{(i)}\right)$ $(j = 1, ..., m)$, with the proposed density estimate for $[U_s]$ having form of (3.9).

3.2.4 Relationship between Gibbs sampling and substitution sampling

It is apparent that in the case of two random variables Gibbs sampling and substitution sampling are identical. For more than two variables, using (3.10) and its obvious generalization to k variables, we see that Gibbs sampling assumes the availability of the set of k full conditional distributions (the minimal set needed to determine the joint density uniquely). The substitution-sampling algorithm requires the availability of $k(k-1)$ conditional distributions, including all of the full conditionals.

Gibbs sampling is known to converge slowly in applications with k very large. Regardless, fair comparison with substitution sampling, in the sense of the total amount of random variate generation, requires that we allow the Gibbs sampling algorithm $i(k-1)$ iterations if the substitution-sampling algorithm is allowed i. even so, there is clearly scope for accelerated convergence for the substitution-sampling algorithm, since it samples from the correct distribution

each time, whereas Gibbs sampling only samples from the full conditional distributions. To amplify, we describe how the substitution-sampling algorithm might be carried out under availability of just the set of full conditional distributions. We see that it can be viewed as the Gibbs sampler, but under an io visiting scheme different from the natural one. We present the argument in the three-variable case for simplicity. Returning to (3.10), if $[X \mid Y]$ is unavailable we can create a sub-substitution loop to obtain it by means of

$$[Y \mid X] = \int [Y \mid X, Z] * [Z \mid X] \tag{3.12}$$

$$[Z \mid X] = \int [Z \mid X, Y] * [Y \mid X]. \tag{3.13}$$

Similar subloops are clearly available to create $[X \mid Z]$ and $[Z \mid Y]$. In fact, for k variables this idea can be straightforwardly extended to the estimation of an arbitrary reduced conditional distribution, given the full conditionals. We omit the details.

The previous analysis suggests that we could view the reduced conditional densities such as $[Y \mid X]$ as available, and that we could thus carry out the substitution algorithm as if all needed conditional distributions were available; however, $[Y \mid X]$ is not available in our earlier sense. Under the subloop in (3.12), we can always obtain a density estimate for $[Y \mid X]$, given any specified X, say $X^{(0)}$. At the next cycle of the iteration, however, we would need a brand-new density estimate for $[Y \mid X]$ at $X = X^{(1)}$. Nonetheless, suppose we persevered in this manner, making our way through one cycle of (3.10). The reader may verify that the only distributions actually sampled from are, of course, the available full conditionals, that at the end of the cycle each full conditional will have been sampled from at least once, and thus that under repeated iterations each variable will be visited io. Therefore, this version of the substitution-sampling algorithm is merely Gibbs sampling with a different but still io visiting order. As a result, *GG1*, *GG2* and *GG3* still hold (*TW1*, *TW2* and *TW3* apply directly only when all required conditional distributions are available). Moreover, there is no gain in implementing the Gibbs sampler

in this complicated order; the natural order is simpler and equally good.

This discussion may be readily extended to the case of k variables. As a result, we conclude that when only the set of k full conditionals is available the substitution-sampling algorithm and the Gibbs sampler are equivalent. Furthermore, we can now see when substitution sampling offers the possibility of acceleration relative to Gibbs sampling. This occurs when some reduced conditional distributions, distinct from the full conditional distributions, are available. Suppose that we write the substitution algorithm with appropriate conditioning to capture these available reduced conditionals. As we traverse a cycle, we would sample from these distributions as we come to them, otherwise sampling from the full conditional distributions.

3.2.5 The Rubin Importance Sampling Algorithm

Rubin[38] suggested a noniterative Monte Carlo method for generating marginal distributions using importance-sampling ideas. We first present the basic idea in the two-variable case. Suppose that we seek the marginal distribution of X, given only the functional form (modulo the normalizing constant) of the joint density $[X, Y]$ and the availability of the conditional distribution $[X \mid Y]$. Suppose further (as is typically the case in applications) that the marginal distribution of Y is not known. Choose an importance-sampling distribution for Y that has density $[Y]_s$. Then, $[X \mid Y] * [Y]_s$ provides an importance sampling distribution for (X, Y). Suppose that we draw iid pairs (X_l, Y_l) $(l = 1, \ldots, N)$ from this distribution, for example, by drawing Y_l from $[Y]_s$ and X_l from $[X \mid Y_l]$. Rubin's idea is to calculate $r_l = [X_l, Y_l] / ([X_l \mid Y_l] * [Y_l]_s)$ $(l = 1, \ldots, N)$ and then estimate the marginal density for $[X]$ by

$$\widehat{[X]} = \frac{\sum_{l=1}^{N} [X \mid Y_l] r_l}{\sum_{l=1}^{N} r_l} \tag{3.14}$$

Note the important fact that $[X, Y]$ need only be specified up to a constant, since the latter cancels in (3.14). In other words, we do not need to evaluate the normalizing constant for $[X, Y]$. By dividing the numerator and the denominator of (3.14) by N and using the law of large numbers, we immediately have the following.

R1 (convergence). $[\widehat{X}] \to [X]$ with probability 1 as $N \to \infty$ for almost every X.

In addition, if $[Y \mid X]$ is available we immediately have an estimate for the marginal distribution of Y :

$$\left[\widehat{Y}\right] = \frac{\sum_{l=1}^{N} [Y \mid X_l]\, r_l}{\sum_{l=1}^{N} r_l}.$$

The successful performance of the density estimator (3.14) depends strongly on the choice of $[Y]_s$ and its closeness to $[Y]$. Thus the suggestion of Tanner and Wong[44] in their rejoinder to Rubin, to perhaps use for $[Y]_s$ the density estimate created after i iterations of the substitution algorithm, merits further investigation.

The extension of the Rubin importance-sampling idea to the case of k variables is clear. For instance, when $k = 3$, suppose that we seek the marginal distribution of X, given the functional form of $[X, Y, Z]$ up to a constant and the availability of the full conditional $[X \mid Y, Z]$. In this case, the pair (Y, Z) plays the role of Y in the two variable case discussed before, and in general we need to specify an importance-sampling distribution $[Y, Z]_s$. Nevertheless, if $[Y \mid Z]$ is available, for example, we only need to specify $[Z]_s$. In any case, we draw iid triples (X_l, Y_l, Z_l) $(l = 1, \ldots, N)$ and calculate $r_l = [X_l, Y_l, Z_l] \,/\, ([X_l \mid Y_l, Z_l] * [Y_l, Z_l]_s)$. The marginal density estimate for $[X]$ then becomes [analogous to (3.14)]

$$\widehat{[X]} = \frac{\sum_{l=1}^{N} [X \mid Y_l, Z_l] r_l}{\sum_{l=1}^{N} r_l} \tag{3.15}$$

We note that in the k-variable case the Rubin importance-sampling algorithm requires Nk random variate generations, whereas Gibbs sampling stopped at iteration i requires mik generations. For fair comparison of the two algorithms, we should therefore set $N = mi$. The relative relationship between the estimators (3.7) and (3.14) may be clarified if resample $Y_1^*, Y_2^*, \ldots, Y_m^*$ from the distribution that places mass $r_l / \sum r_l$ at Y_l $(l = 1, \ldots, N)$. we could then replace (3.14) with

$$\widehat{[X]} = \frac{1}{m} \sum_{j=1}^{m} \left[X \mid Y_j^*\right], \tag{3.16}$$

So (3.7) and (3.16) are of the same form. Relative performance on average depends on whether the distribution of $Y^{(i)}$ or Y^* is closer to $[Y]$. Empirical work described in Gelfand and Smith[15] suggested that under fair comparison (3.7) performs better than (3.15) or (3.16). It seems preferable to iterate through a learning process with small samples rather than to draw a one-off large sample at the beginning.

3.3 Density Estimation

In this section, we consider the problem of calculating a final form of marginal density from the final sample produced by either the substitution or Gibbs sampling algorithms. Since for any estimated marginal the corresponding full conditional has been assumed available, efficient inference about the marginal should clearly be based on using this full conditional distribution. The next section presents several methods for dealing with unavailable conditional distributions.

In the simplest case of two variables, this implies that $[X \mid Y]$ and the $Y^{(j)}$ $(j = 1, \ldots, m)$ should be used to make inferences about $[X]$, rather than imputing $X_j^{(i)}$'s. Intuitively, this follows, because to estimate $[X]$ using the $X_j^{(i)}$ requires a density estimate. Silverman[42] provides many techniques to construct a density estimate for $[X]$ based on $X_j^{(i)}$'s. This estimate can be adequate if at the last iteration the number of replications, m, is large enough. However, such an estimate ignores the known form of $[X \mid Y]$ that is mixed to obtain $[X]$. This alternative density estimate, having the form of (3.9), is better under a wide range of loss functions. The formal argument is essentially based on the Rao-Blackwell theorem. Gelfand and Smith[15] sketched a proof in the context of density estimator.

If X is a continuous p-dimensional random variable, consider any kernel density estimator of $[X]$ based on the $X_j^{(i)}$ (e.g., see Devroye and Györfi[8]) evaluated at X_0 : $\Delta_{X_0}^{(i)} = (1/mh_m^p)\sum_{j=1}^{m} K[(X_0 - X_j^{(i)})/h_m]$, say, where K is a bounded density on R^p and the sequence $\{h_m\}$ is such that as $m \to \infty$, $h_m \to 0$, whereas $mh_m \to \infty$. To simplify notation, set $Q_{m,X_0}(X) = (1/h_m^p)K[(X_0 - X)/h_m]$ so that $\Delta_{X_0}^{(i)} = (1/m)\sum_{j=1}^{m} Q_{m,X_0}(X_j^{(i)})]$. Define $\gamma_{X_0}^{(i)} = (1/m)\sum_{j=1}^{m} E(Q_{m,X_0}(X_j^{(i)}) \mid Y_j^{(l)}]$. By our earlier theory, both $\Delta_{X_0}^{(i)}$ and $\gamma_{X_0}^{(i)}$ have the same expectation. By the Rao-Blackwell theorem:

$varE\left(Q_{m,X_0}\left(X \mid Y\right)\right) \leq varQ_{m,X_0}(X)$, and hence $MSE\left(\gamma_{X_0}^{(i)}\right) \leq MSE\left(\Delta_{X_0}^{(i)}\right)$, where MSE denotes the mean squared error of the estimate of $[X_0]$.

Now, for fixed Y, as $m \to \infty$, $E\left(Q_{m,X_0}\left(X \mid Y\right)\right) \to [X_0 \mid Y]$ for almost every X_0, by the Lebesgue density theorem (see Devroye and Györfi[8], p.3). Thus in terms of random variables we have $E\left(Q_{m,X_0}\left(X \mid Y\right)\right) \xrightarrow{d} [X_0 \mid Y]$, so for large m, $\gamma_{X_0}^{(i)} \sim \left[\widehat{X_0}\right]_i$ and $MSE\left(\gamma_{X_0}^{(i)}\right) \approx MSE\left(\left[\widehat{X_0}\right]_i\right)$, and hence $\left[\widehat{X_0}\right]_i$ is preferred to $\Delta_{X_0}^{(i)}$.

The argument is simpler for estimation of $\eta = E\left(T\left(X\right)\right) = \int T\left(X\right) * [X]$, say. Here, $\widehat{\eta}_1 = (1/m)\sum_{j=1}^{m} T\left(X_j^{(i)}\right)$ is immediately seen to be dominated by $\widehat{\eta}_2 = (1/m)\sum_{j=1}^{m} T\left(X_j^{(i)} \mid Y_j^{(i)}\right)$.

3.4 Sampling Issues

Basic to the implementation of the Gibbs sampler is the ability to sample from the conditional distribution $[X_i \mid X_j, j \neq i]$. Carlin and Gelfand[3] referred to this property as *conjugacy.* In the terminology of Tanner and Wong[44] it is assumed that one can sample directly from the *augmented posterior distribution.* Techniques for the efficient generation of appropriate random variates from conjugate conditionals are described in detail in Devroye[7] and Ripley[32]. Gelfand, Hills, Racine-Poon, and Smith[14] presented various examples that are conjugate. but other classes of problems exist (e.g. nonlinear regression) where the posterior distribution is lacking conjugacy in at least one of the conditionals.

Recently, several methods have been proposed for dealing with nonconjugate conditionals via importance sampling or acceptance/rejection approaches. Wei and Tanner[47] presented an importance sampling modification when this simplicity is absent. Carlin and Gelfand[3] and Zeger and Karim[49] presented rejection/acceptance modifications to the Gibbs sampler. Also of note is the work by Gilks and Wild[18], who used tangent and secant approximations above and below the log-posterior to develop an acceptance/rejection scheme. Acceptance/rejection methods are exact in the sense that they produce samples from the required distribution. Possible drawbacks to theses methods include a low acceptance rate, the need to know the normalizing constant for the conditional distribution, and possible restrictions on the distribution (e.g., log-concavity). Nonlinear regression problems for example, may lead to conditionals that are not log-concave and may in fact be multimodal. Moreover, in nonlinear regression the conditionals typically are known only up to a multiplicative constant. Both importance sampling and acceptance/rejection algorithms require a higher degree of programming sophistication on the part of the data analyst and thereby detract from the conceptual simplicity and appeal of the Gibbs sampler. Specification of an importance sampling function that is easy to sample from, yet provides a "good match" to the density of interest, may require the specification of several tuning constants. In the context

of rejection/acceptance, specification of tuning constants is required to ensure that the importance function, modulo a multiplicative constant, dominates the density of interest *everywhere* (or at least over a region of high content). Several sophisticated and more efficient variants of the method have been developed. A fast and effective algorithm for unimodal densities is given in Zaman[48].

In many applications of the Gibbs sampler (for instance, Carlin and Gelfand[3]; Carlin et al.[4]; Gelfand and Smith[15]; Gelfand et al.[14]; Gelfand et al.[16]; Zeger and Karim[49]), the conditional distribution $[X_i \mid X_j, j \neq i]$ is univariate. Ritter and Tanner[33] used this observation to develop an approach that they called the *Griddy-Gibbs sampler*, for sampling from the conditional distribution in the absence of conjugacy, which preserves the conceptual and implementational simplicity of the Gibbs sampler.

3.4.1 The Griddy-Gibbs Sampler

As noted previously, in many practical situations the distribution $[X_i \mid X_j, j \neq i]$ is univariate. When it is difficult to sample directly from it, the idea is to form a simple approximation to the inverse cdf based on the evaluation of $[X_i \mid X_j, j \neq i]$ on a grid of points. More formally perform the following steps:

Step 1. Evaluate $[X_i \mid X_j, j \neq i]$ at $X_i = x_1, \ldots, x_n$ to obtain $w_1, \ldots, w_n$.

Step 2. Use $w_1, \ldots, w_n$ to obtain an approximation to the inverse cdf of $[X_i \mid X_j,\ j \neq i]$.

Step 3. Sample a uniform (0,1) deviate and transform the observation via the approximate inverse cdf.

Remark 1 *The function* $[X_i \mid X_j, j \neq i]$ *need be known only up to a multiplicative constant, because the normalization can be obtained directly from the* $w_1, \ldots, w_n$.

Remark 2 *The grid $x_1, \ldots, x_n$ need not be uniformly spaced. In fact, good grids put more points in neighborhoods of high mass and fewer points in neighborhoods of low mass. The approach of Ritter and Tanner[33] to address this goal is to construct the grid so that the mass under the current approximation to the conditional distribution between successive grid points is approximately constant.*

Remark 3 *The number of points in the grid need not be constant over the iterations of the Gibbs sampler. At early iterations, n may be small. As the algorithm iterates toward the joint distribution, n can be increased.*

Remark 4 *Simple approximations to the inverse cdf are:*

(a) *Piecewise constant corresponding to a discrete distribution for $x_1, \ldots, x_n$, with probabilities $p(x_i) = w_i / \sum_{j=1}^{n} w_j$.*

(b) *Piecewise linear corresponding to a piecewise uniform distribution on the interval $[a_i, a_{i+1}]$, $i = 1, \ldots, n$, with $x_i \in [a_i, a_{i+1}]$ and density $f_i = w_i / \sum_{j=1}^{n} \omega_j$, where $\omega_i = w_i(a_{i+1} - a_i)$. Typically, x_i is centered in the interval $[a_i, a_{i+1}]$.*

More sophisticated approximations may be based on piecewise quadratic interpolation or using other types of splines. Many such methods are available in such libraries as IMSL. In general, when the conditional is easy to evaluate, one may wish to use a simpler approximation to the inverse cdf and use a finer grid. When the conditional is more difficult to evaluate, one may wish to use a coarser grid but a more clever approximation.

3.5 Convergence Issues

Complete implementation of the Gibbs sampler or substitution algorithm requires a determination of i be made and that, across iterations, choice(s) of

m be specified. In this regard it is important to distinguish the assessment of convergence for any individual data application from the broader goal of developing on-line, automated, interactive software to determine satisfactory convergence. Raftery and Lewis[30] focused on quantiles of functionals of the posterior distribution, and described an easily implemented method for determining the total number of iterations required, and also the number of initial iterations that should be discarded. Roberts[35] addressed the issue of convergence and its diagnosis. He focused on searching for a one-dimensional summary statistic, which is calculated in each iteration, and which attempts to describe the convergence mechanism. Chan[6] studied the asymptotic behavior of the Gibbs sampler. He obtained mild sufficient conditions for the ergodicity of the Gibbs sampler and discussed its geometric ergodicity. Roberts and Smith[36] presented simple conditions which ensure the convergence of the Gibbs sampler. Robert[34] presented some convergence control methods for Markov Chain Monte Carlo algorithms. However, the above approaches showed little use in applications. Simpler methods to be discussed in more detail are more common to monitor convergence during the implementation of the Gibbs sampler or the data augmentation algorithm. Gelfand, Hills, Racine-Poon and Smith[14] gave a brief discussion on this issue based on their extensive experience with a wide range of applications. Tanner and Wong[44] suggested some graphical methods to monitor the progress of the data augmentation algorithm. And finally, Ritter and Tanner[33] discussed an importance sampling technique that they called the *Gibbs stopper* for assessing convergence of the Gibbs sampler for moderate sized problems. This technique is also useful for converting the output of the Gibbs sampler to a sample from the exact posterior.

3.5.1 "Gelfand, Hills, Racine-Poon and Smith" methods

The extensive experience of Gelfand, Hills, Racine-Poon and Smith with a wide range of particular applications suggests that assessing the convergence of the Gibbs sampler is not a problem. They note that appropriate values for i and

m depend upon the particular application and cannot be specified in advance. All of the examples discussed in Gelfand, Hills, Racine-Poon and Smith[14] were handled with $i \leq 50$ and $m \leq 1000$. Since random variate generation is generally inexpensive, we expect to experiment with different settings. Indeed, since interest focuses heavily on the application of sampling-based approaches to previously inaccessible problems where we often have no benchmarks or alternatives with which to compare our results, such experimentation seems necessary.

The following discussion describes a means of assessing convergence that, though naive and less rigorously defined than might be desired, has been successful in a considerable number of applications. We monitor the generated data in a univariate fashion, allowing the sampler to run until we feel that the marginal posterior distributions for each parameter of interest are converged. We do this in an elementary manner. For a fixed m we increase i, overlay plots of the resulting estimated densities (3.9), and see if the estimates are visually indistinguishable. Similarly, we also increase m to assess stability of the density estimate. We tend to hold m somewhat small (often as small as 25 and at most 200) until convergence is indicated, at which point for a final iteration, we typically increase m by an order of magnitude to obtain our density estimate (3.9). Univariate plots are drawn by selecting between 40 and 100 equally spaced points in the effective domain of the variable. We then evaluate the density estimate at these points and a spline-smoothed curve is drawn through these values. By effective domain we mean the interval where, say, 99% of the mass lies. Clearly, this plotting method could be refined. In this regard, we also recommend a convenient check on calculations by performing a simple trapezoidal integration on the collection of estimated density values associated with these points to verify that the result is very close to 1.

3.5.2 Tanner and Wong methods

Tanner and Wong[44] pointed out that it is helpful to graphically monitor the progress of the data augmentation algorithm, for example, using selected

percentiles of the estimated posterior distribution. If one is interested in first and second moments , then these moments, rather than extreme tail behavior, may be monitored. The data augmentation algorithm (for a fixed value of m) may be iterated until the fluctuations in such a plot indicate that the process has become stationary. At such a point, the algorithm may be terminated or the value of m increased to improve the precision (with respect to Monte Carlo variation) of the estimate of the functional of the posterior of interest. In this way, start with the smaller value of m and then increase the value of m at various junctures in the iteration process to realize computational savings.

3.5.3 The Gibbs Stopper

The idea is to assign the weight w to the d-dimensional vector X that has been drawn from the current approximation to the joint distribution g_i. This technique was mentioned in the rejoinder of Tanner and Wong[44] and was implemented in Wei and Tanner[47] in the context of data augmentation. The appropriate weight is given as

$$w = \frac{q(X)}{g_i(X)},$$

Where $q(X)$ is proportional to $p(X)$. By carrying along these weights, one realizes a sample from $p(X)$ rather than from the approximation. But because the variance of any function of these tuples will depend on $p(X)^2/g_i(X)$, it is important to iterate the Gibbs sampler to help eliminate outlying weights that would inflate the Monte Carlo variance of the estimate.

To write down the functional form for g_i, we introduce notation following Schervish and Carlin[39]. Let $p^{(i)}(X) = p(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$. For two vectors X and X', define for each $i = 1, \ldots, d-1$,

$X^{(i')} = \left(X_1, \ldots, X_{i-1}, X'_{i+1}, \ldots, X'_d\right)$. We adopt the convention that $X^{(0')} = X'$, $X^{(d')} = X$, and $p^{(0)}(X) = p(X)$.

As noted in Schervish and Carlin[39], if g_i is the joint distribution of the observations sampled at iteration i, then the joint density (g_{i+1}) of the observations sampled at the next iteration is given by

$$\int K(X', X)\, g_i(X')\; d\lambda(X'), \tag{3.17}$$

Where

$$K(X', X) = \prod_{i=1}^{d} p^{(i)}\left(X^{(i')}\right).$$

One may approximate (3.17) via the method of Monte Carlo. In particular, given the observations $X^1, \ldots, X^m$ sampled at iteration i, use the Monte Carlo sum $\frac{1}{m}\sum_{j=1}^{m} K(X^j, X)$ to approximate $g_{i+1}(X)$.

It is noted that if the current approximation to the joint distribution is "close" to the joint distribution, then the distribution of the weights will be degenerate about a constant. In this way one could construct a serious of plots, each of which presents the distribution of the weights (over replicated paths of the Gibbs sampler) at a given iteration. A numerical approach would consist of computing some functional of the distribution of the weights (e.g., interquartile range or standard deviation) and then monitor this value as the iterations increase. As the algorithm is iterated, this functional should begin to fluctuate about a value and the weight plots should move toward a spike distribution.

3.6 Gibbs Sampling with Improper Posteriors

If a complicated hierarchical model with improper priors is postulated, then it will be often the case that demonstration of propriety of the posterior will be mathematically tedious, if not impossible. On the other hand, many such models have some type of conjugate structure that makes calculation of the Gibbs

conditionals a simple exercise in the recognition of common functional forms. However, if propriety has not been demonstrated, then calculating these densities via recognition requires assuming (possibly incorrectly) that a proportional form of the joint density holds. If a set of proper densities results, then it is tempting to assume that the posterior distribution is proper–but this may not be true. The end result is that the Gibbs sampler may be used in conjunction with a set of conditionals corresponding to an improper posterior distribution. Gelfand et al.[14] and Wang et. al[46] analyzed data using the Gibbs sampler in conjunction with one-way random-effects models with improper posteriors. Both of this articles showed plots of approximate marginal posterior densities and gave other results which seem completely reasonable. It is not at all obvious in theses examples that the posterior distribution is improper and, in fact, that all inferences are to nonexistent posterior distributions.

Hobert and Casella[25] showed that Gibbs Markov chains constructed using conditionals from an improper posterior are null (i.e., null recurrent or transient) and thus do not enjoy the convergence properties associated with Gibbs chains corresponding to proper posteriors. Thus, although Monte Carlo approximations based on the output from a null Gibbs chain may appear reasonable (as in Gelfand et al.[14] and Wang et al.[46]), their limiting behavior is often quite unreasonable (Hobert and Casella[24]). One way to avoid improper posteriors is to use proper priors. In mixed linear models, ignorance can be modeled by using a normal prior with very large variance for the fixed effects and inverted gamma priors with very small parameter values for the variance components.

Hobert[23] experimented with some diagnostics for null Markov chains but had not been met with much success. Typically, the diagnostics work well only in cases where they are not really needed; that is, when the Markov chain is clearly misbehaved. Note that the common diagnostics for Monitoring convergence of the Markov chain are not really appropriate for the cases of improper posteriors. These diagnostics are working under the assumption that the Markov chain is positive recurrent. Thus they are not diagnosing if the chain will converge, but rather when it will converge. It seems that, for now,

the only foolproof way of avoiding the problem is to use proper priors or results like those of Dey, Gelfand, and Peng[9] and Ibrahim and Laud[26], which give sufficient conditions for propriety of posteriors for classes of improper priors.

3.7 Summary Discussion

We have emphasized providing a comparative review and explication of three possible sampling approaches to the calculation of intractable marginal densities. The substitution, Gibbs, and importance-sampling algorithms are all straightforward to implement in several frequently occurring practical situations, thus avoiding complicated numerical or analytic approximation exercises. For this latter reason if not for others the techniques deserve to be better known and experimented with a wide range of problems. We also provided some important tools in sampling and assessing convergence to facilitate the implementation of these algorithms.

We hope that the unified exposition attempted here will provide a general, clarifying perspective within which to view the work of Geman and Geman[17], Rubin[38, 37], and Tanner and Wong[44], and to evaluate its potential for the Hierarchical Bayes problem discussed in Chapter 2. In Chapter 4, we will illustrate how we can implement two of the above algorithms; the Gibbs sampler and the importance-sampling algorithm; on the QMP model.

Chapter 4

QMP using Monte Carlo Methods

4.1 The Hierarchical Bayes Model

4.1.1 Model 1

Let us formulate the QMP model in HB terms and add elements necessary to complete the specification. Suppose that we observe independent number of defects, x_t, over different rating periods $t = 1, \ldots, T$ having expected number of defects e_t (with resultant sample quality index $I_t = x_t/e_t$). We assume $[x_t \mid \theta_t] = Poisson(e_t\theta_t)$, and θ_t are iid from $Gamma(\alpha, \beta)$, with density $\theta_t^{\alpha-1}e^{-\theta_t e_t}/\beta^\alpha\Gamma(\alpha)$. For model 1, let us assume that we have flat priors on the hyperparameters α and β. That is to say, the prior information on α and β is modeled with the improper densities $\rho(\alpha) = \rho(\beta) = 1$.

To implement the Gibbs sampler, we need three conditional distributions. The conditional distribution for θ_t, the conditional distribution for β and the conditional distribution for α. These are given in the following lemma.

Lemma 1 *In the HB model specified above, the conditional distribution of*

1. $[\theta_t \mid \mathbf{x}, \beta, \alpha, \theta_{s,s\neq t}] = G\left(\alpha + x_t, (1/\beta + e_t)^{-1}\right), t = 1, \ldots, T.$

2. $[\beta \mid \mathbf{x}, \alpha, \theta_1, \ldots, \theta_T] = IG\left(T\alpha - 1, \sum \theta_t\right).$

3. $[\alpha \mid \mathbf{x}, \beta, \theta_1, \ldots, \theta_T] = k\frac{\left(\left(\prod_{t=1}^{T} \theta_t\right)/\beta^T\right)^{\alpha}}{\Gamma(\alpha)^T}$, *where* k *is the normalizing constant.*

Proof. The conditional distribution of θ_t was already computed in Section 2.3.1.

The conditional distribution of β can easily be derived as follows:

$$\begin{aligned}
[\beta \mid \mathbf{x}, \alpha, \theta_1, \ldots, \theta_T] &\propto [\beta, \alpha, \mathbf{x}, \theta_1, \ldots, \theta_T] \\
&\propto \left(\prod_{t=1}^{T} [\theta_t \mid \beta]\right) \times [\beta] \\
&= \left(\prod_{t=1}^{T} \frac{\theta_t^{\alpha-1} e^{-\theta_t/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\right) \times 1 \\
&\propto \left(\frac{e^{-\sum \theta_t/\beta}}{\beta^{\alpha T}}\right) \\
&= \frac{e^{-\sum \theta_t/\beta}}{\beta^{(\alpha T-1)+1}}. \qquad (4.1)
\end{aligned}$$

From (4.1), we recognize β given $\mathbf{x}, \alpha, \theta_1, \ldots, \theta_T$ to be distributed as an inverse gamma distribution $IG\left(T\alpha - 1, \sum \theta_t\right)$.

The conditional distribution of α can easily be derived as follows:

$$\begin{aligned}
[\alpha \mid \mathbf{x}, \beta, \theta_1, \ldots, \theta_T] &\propto [\alpha, \beta, \mathbf{x}, \theta_1, \ldots, \theta_T] \\
&\propto \left(\prod_{t=1}^{T} [\theta_t \mid \alpha]\right) \times [\alpha] \\
&= \left(\prod_{t=1}^{T} \frac{\theta_t^{\alpha-1} e^{-\theta_t/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\right) \times 1 \\
&\propto \frac{\left(\left(\prod_{t=1}^{T} \theta_t\right)/\beta^T\right)^{\alpha}}{\Gamma(\alpha)^T}. \qquad (4.2)
\end{aligned}$$

From (4.2), we recognize that the conditional distribution of α is not a known density. It is known just up to a multiplicative constant. Its normalizing constant k can not be calculated analytically. ■

Prior to the implementation of the Monte Carlo approaches discussed in the previous chapter, we should discuss some implementation issues that might affect their efficiency. We notice in the above lemma that the conditional distribution of α lacks conjugacy as explained in Section 3.4. This problem can be solved by the techniques discussed in that section. However, implementing these techniques reduced considerably the time efficiency of the Gibbs sampler approach. Knowing that the QMP algorithm is quite rapid, time efficiency is an important factor in developing an algorithm that might perform better than the QMP algorithm.

Another problem with model 1, is that the use of improper priors on α and β may not be plausible. These forms of uninformative priors are mostly used in applications where not much prior information is available about the hyperparameters. However, in our case there is much information on which we can improve our estimation procedures.

Considering all the issues discussed above, we formulate a simpler model that improves the efficiency of the Gibbs sampler and use the prior information available in the system.

4.1.2 Model 2

Let us reformulate model 1, and add the necessary elements to a build a new HB model. As specified in model 1, we assume $[x_t \mid \theta_t] = Poisson\,(e_t\theta_t)$,and θ_t are iid from $Gamma\,(\alpha, \beta)$. The parameter α can be considered as a tuning parameter and can be estimated using the marginal distribution of the x_t's or the method-of-moments as shown later. Moreover, it is plausible to assume that β arises from an inverse gamma distribution $IG(\gamma, \delta)$ with density $\delta^{\gamma}e^{-\delta/\beta}/\beta^{\gamma+1}\Gamma(\gamma)$. This is due to the fact that assuming a flat prior (as in

model 1), the conditional distribution of β is an inverted gamma. A diffuse version of this final stage distribution is obtained by taking γ and δ to be very small. This completes the specification of the Bayesian model. The posterior distributions $[\theta_t \mid \mathbf{x}]$ are sought.

One possible estimator for α is $\widehat{\alpha}_1 = \overline{I}^2/(S_T^2 - T^{-1}\overline{I}\sum_{t=1}^{T} e_t^{-1})$, with the latter derived by the method-of-moments empirical Bayes argument based on:

$$\begin{aligned} E(I_t) &= EE(I_t \mid \theta_t) = \alpha\beta \approx \overline{I}. && (4.3) \\ var(I_t) &= varE(I_t \mid \theta_t) + Evar(I_t \mid \theta_t) \\ &= \alpha\beta/e_t + \alpha\beta^2 \approx S_T^2 = T^{-1}\sum_{t=1}^{T}\left(I_t - \overline{I}\right)^2. && (4.4) \end{aligned}$$

The first problem with $\widehat{\alpha}_1$ as an estimate of α is that it can be negative. To solve this problem, we just use (4.3) to obtain $\widehat{\alpha}_2^{(i)} = \overline{I}/\beta^{(i)}$, where $\beta^{(i)}$ is an estimator of β at iteration i of the Gibbs sampler algorithm. So $\widehat{\alpha}_2^{(i)}$ will approximately converge to α as $\beta^{(i)}$ converges to β. The second problem that might arise with $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ is the case where $I_t = 0$ for all t. This will make $\widehat{\theta}_t = 0$. But $\widehat{\theta}_t$ is a posterior mean of a positive parameter, so it cannot be zero. The correct method of handling this problem is to start with a proper distribution on α. But then the mathematics and the computations become complicated. So we assert that we have prior information that is equivalent to observing some prior data x_0 and e_0.

After specifying how to estimate α, we can now consider it as known and derive as in model 1, the conditional distribution for θ_t and the conditional distribution for β. These are given in the following lemma.

Lemma 2 *In the HB model specified above, the conditional distribution of*

1. θ_t *given* $\mathbf{x}, \beta, \theta_{s,s\neq t}$ *is* $G\left(\alpha + x_t, (1/\beta + e_t)^{-1}\right)$, $t = 1, \ldots, T$.

2. β *given* $\mathbf{x}, \theta_1, \ldots, \theta_T$ *is* $IG\left(\gamma + T\alpha, \sum\theta_t + \delta\right)$.

Proof. The conditional distribution of θ_t was already computed in section 2.3.1. The conditional distribution of β can easily be derived as follows:

$$\begin{aligned}
[\beta \mid \mathbf{x}, \theta_1, \ldots, \theta_T] &\propto [\beta, \mathbf{x}, \theta_1, \ldots, \theta_T] \\
&\propto \left(\prod_{t=1}^{T} [\theta_t \mid \beta] \right) \times [\beta] \\
&= \left(\prod_{t=1}^{T} \frac{\theta_t^{\alpha-1} e^{-\theta_t/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \right) \times \left(\frac{\delta^{\gamma} e^{-\delta/\beta}}{\beta^{\gamma+1}\Gamma(\gamma)} \right) \\
&\propto \left(\frac{e^{-\sum \theta_t/\beta}}{\beta^{\alpha T}} \right) \times \left(\frac{e^{-\delta/\beta}}{\beta^{\gamma+1}} \right) \\
&= \frac{e^{-\frac{\left(\delta + \sum \theta_t\right)}{\beta}}}{\beta^{\alpha T+\gamma+1}} \qquad (4.5)
\end{aligned}$$

From (4.5), we recognize β given $\mathbf{x}, \theta_1, \ldots, \theta_T$ to be distributed as an inverse gamma distribution $IG(\gamma + T\alpha, \sum \theta_t + \delta)$. ■

Now, we have specified in Lemma 2 the required conditional densities needed to implement the Gibbs sampler on model 2. We notice that we have just two conditional densities, so the substitution and Gibbs sampler algorithm are equivalent as we have showed in Section 3.2.4. However, the specifications are little different for the importance-sampling algorithm as we will show in the next section.

4.2 Monte Carlo Algorithms

4.2.1 Gibbs Sampler Algorithm

We can now apply the Gibbs sampler algorithm on model 2. Given $\left(\theta_1^{(0)}, \ldots, \theta_T^{(0)}, \beta^{(0)}\right)$, the Gibbs sampler draws $\theta_t^{(1)} \sim G\left(\alpha + x_t, (1/\beta^{(0)} + e_t)^{-1}\right)$, $(t = 1, \ldots, T)$ and $\beta^{(1)} \sim IG\left(\gamma + T\alpha, \sum \theta_t^{(1)} + \delta\right)$ to complete one cycle. For estimating α, we can use $\widehat{\alpha}_1$ if ever it is positive, or use $\widehat{\alpha}_2$ as discussed in the previous section. Now,

If we carry out m repetitions each of i iterations, generating $\left(\theta_{1l}^{(i)}, \ldots, \theta_{Tl}^{(i)}, \beta_l^{(i)}\right)$ $(l = 1, \ldots, m)$, the posterior density estimate for θ_t at each iteration is

$$[\theta_t \uparrow \mathbf{x}]_i = \frac{1}{m} \sum_{l=1}^{m} G\left(\alpha + x_t, (1/\beta_l^{(i)} + e_t)^{-1}\right), \; t = 1, \ldots, T, \tag{4.6}$$

whereas

$$[\beta \uparrow \mathbf{x}]_i = \frac{1}{m} \sum_{l=1}^{m} IG\left(\gamma + T\alpha, \sum \theta_{tl}^{(i)} + \delta\right). \tag{4.7}$$

From the posterior density estimate of θ_t given in (4.6), we can obtain point estimates of θ_t's at each iteration as follows:

$$\widehat{\theta}_t^{(i)} = \frac{1}{m} \sum_{l=1}^{m} \frac{\alpha + x_t}{1/\beta_l^{(i)} + e_t}, \; t = 1, \ldots, T.$$

For future reference, we will denote this estimator as the *Gibbs estimator.*

4.2.2 Importance-Sampling Algorithm

The estimator of the marginal density of θ_t under the Rubin's importance sampling algorithm is

$$[\theta_t \uparrow \mathbf{x}] = \frac{\sum_{l=1}^{N} G\left(\alpha + x_t, (1/\beta_l + e_t)^{-1}\right) \times r_l}{\sum_{l=1}^{N} r_l}, \; t = 1, \ldots, T. \tag{4.8}$$

Here $r_l = [\mathbf{x} \mid \beta_l] * [\beta_l] / [\beta_l \mid \mathbf{x}]_s$, where $[\mathbf{x} \mid \beta_l]$ is the product of negative binomial densities as shown in Section 2.3.1; that is ,

$$[\mathbf{x} \mid \beta_l] = \prod_{t=1}^{T} \left(\frac{\Gamma(x_t + \alpha)\, e_t^{x_t} \beta_l^{\alpha}}{x_t! \Gamma(\alpha) (e_t + 1/\beta_l)^{x_t + \alpha}} \right).$$

Period	x_t	e_t
7701	1.19	1.81
7702	2.61	1.76
7703	1.76	1.78
7704	1.92	1.81
7705	1.45	1.80
7706	1.41	1.82

Table 4.1: Equivalent Defects in Keys of Telephone sets - Shreveport

and $[\beta_l]$ is the IG prior evaluated at β_l. A good choice for the importance sampling distribution $[\beta_l \mid \mathbf{x}]_s$ is $IG(\gamma + T\alpha, \sum I_t + \delta)$. This arises because $[\beta_l \mid \mathbf{x}] = E_{[\theta_1,\ldots,\theta_T|\mathbf{x}]}[\beta \mid \mathbf{x}, \theta_1, \ldots, \theta_T] \approx \left[\beta \mid \mathbf{x}, \widehat{\theta}_1, \ldots, \widehat{\theta}_T\right]$, using $\widehat{\theta}_t = I_t$ in the conditional distribution of β given in lemma 2.

From the posterior density estimate of θ_t given in (4.8), we can obtain point estimates of θ_t's as follows:

$$\widehat{\theta}_t = \frac{\sum_{l=1}^{N} \frac{\alpha + x_t}{1/\beta_l + e_t} \times r_l}{\sum_{l=1}^{N} r_l}, \quad t = 1, \ldots, T.$$

For future reference, this estimator will be denoted as the *Rubin's estimator*.

4.2.3 Implementation Issues

We apply model 2 using the Gibbs sampler algorithm discussed above to a part of data on equivalent defects in keys of telephone sets manufactured in Shreveport. This data was previously analyzed by Hoadley (private communication, 1992) as an illustration of the QMP algorithm on a real quality control data set. The data is reproduced here in Table 4.1, where x_t is the equivalent number of defects, and e_t is the expected number of equivalent defects.

We first illustrate the use of the Gibbs sampler to this data set, with $\delta = 1$ and $\gamma = 0.1$. As indicated in chapter 2, if the sample size m is taken to be large in each iteration, then the algorithm can be interpreted as the method of

successive substitution for solving a fixed point problem. In practice, however, it is inefficient to take m large during the first few iterations when the estimated posterior is far from the true distribution. Rather, it is suggested that m initially be small and then increased with successive iterations. In addition, we have found it helpful to monitor the progress of the algorithm by examining the mean of the estimated posterior distribution of θ_t. Moreover, the convergence of the algorithm will also be assessed by graphical displays of the densities at different iterations.

To illustrate these ideas, let us return to our data set. At the initial iteration, m is taken to be 10. The algorithm then runs through 7 iterations, at which it appears (see Fig. 4.1) that the process has become stationary. The sample size is then increased to 100 and the algorithm proceeds through 8 further iterations. The final 5 iterations are run with $m = 200$, and the estimated posterior distribution is then obtained by pooling the values from these 5 iterations. From Figure 4.1, we see that the effect of increasing m has been to reduce substantially the system variability and assess the stability of the density estimate. The estimated posterior for θ_t is obtained from (4.6).

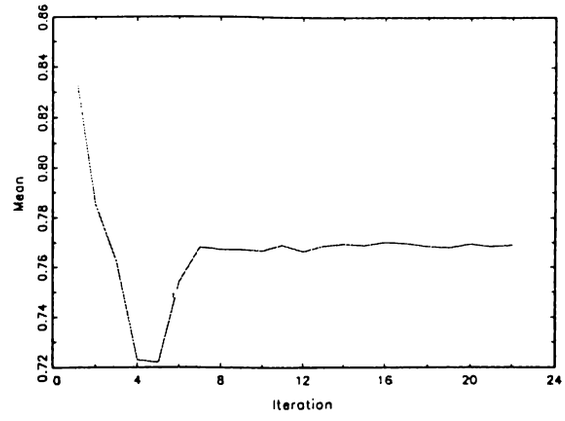


Figure 4.1: The mean of $\widehat{\theta}_6$ across iterations.

Typically, graphical displays, such as Figure 4.1 or any other summary

statistics of the posterior distribution, will give a good idea of how m should be varied and how many iterations are needed. Other important graphical displays that are helpful in assessing the convergence of the estimated posterior, are the densities plots. Figure 4.2, gives the plots of the estimated posterior densities of $\widehat{\theta}_6$ for iterations 5 and 10. The density in iteration 5 (the dashed line) is different from the one in iteration 10 (solid line). This indicates that we still can not assess convergence from just 5 iterations. However, Figure 4.3 shows that the plots of the densities at iterations 10 (solid line) and 15-20 (dashed line) are hardly distinguishable. This is a remarkable convergence from such a small number of drawings.

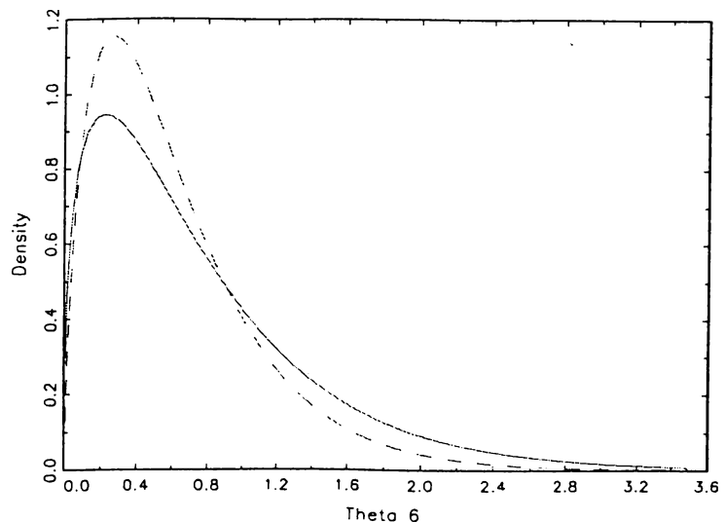


Figure 4.2: The posterior density of $\widehat{\theta}_6$. The dashed line represent iteration 5, and the solid line represent iteration 10.

Finally, it is noted that the previous computations of the Gibbs sampler algorithm required about 1 minute, using "Gauss" on an "IBM 486/33 PC". It is also noted that the convergence properties of the importance-sampling algorithm has not been studied. This algorithm is severely criticized in literature due to its poor performance as explained in Chapter 3. However, in the next section we will use simulation to test the performance of the Rubin's estimator

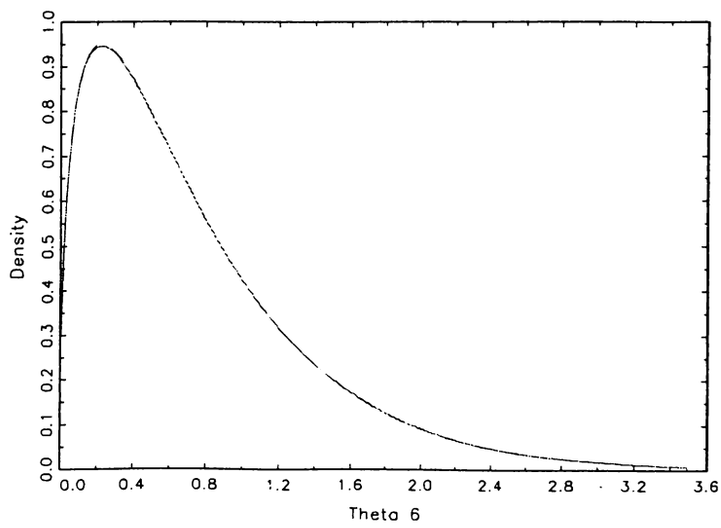


Figure 4.3: The posterior density of $\widehat{\theta}_6$. The dashed line represent iteration 15-20, and the solid line represent iteration 10.

with the Gibbs and QMP estimators, under a fair comparison with the Gibbs sampler (see section 3.2.5).

4.3 Simulation Study

It is of interest to compare the performance of the Gibbs and Rubin's estimator with the QMP estimator under a variety of process distributions. To get a better understanding of how to deal with this, we use simulation.

4.3.1 Simulation Design

Our simulation design is the following:

1. The data size T is equal to 6. We restrict our past data to five periods following the administrative rules of Bell laboratories.

2. The expected number of defects is equal to 5 during all the rating periods.

3. $\theta_1, \ldots, \theta_T$ is a random sample from a process distribution. The eight used are (a) gamma distribution with mean = 1, variance = 1; (b) gamma distribution with mean = 2, variance = 2; (c) point mass at 1; (d) point mass at 1.5; (e) uniform distribution on $[0, 2]$; (f) uniform distribution on $[0.5, 1.5]$; (g) two-point distribution with $\theta_t = 1$ or 2 and $\Pr\{\theta_t = 1\} = 0.75$; (h) two-point distribution with $\theta_t = 0.5$ or 4 and $\Pr\{\theta_t = 0.5\} = 0.5$.

4. The number of iterations i and replications m in the Gibbs sampler are similar to the example of the previous section.

5. The Rubin importance-sampling algorithm is performed under a fair comparison with the Gibbs sampler (i.e. $N = mi$).

There are 8 simulation runs. Each simulation run corresponds to one process distribution. The number of data sets (i.e., simulation iterations) for each simulation run are $J = 50,000$. This number was chosen to make the standard deviations of the estimated percentage of errors (to be defined precisely in the next section) of the various estimators less than 3.

4.3.2 Simulation Results

The results of the simulation are given in Table 4.2. The three columns labeled "Percentage of error ($\overline{p}$)/standard deviation (s_p)" contain two results separated by a slash. To define these results precisely for a given run and estimator, let $\left(\theta_j, \widehat{\theta}_j\right)$, $j = 1, \ldots, J$, denote the true and estimated quality level (of the sixth rating period) for iteration j of the simulation. The percentage of error for iteration j is defined by

Run no.	Process dist.	$\overline{p}/s_p$ for QMP	$\overline{p}/s_p$ for Rubin	$\overline{p}/s_p$ for Gibbs
1	a	32/.8	29/.5	18/.4
2	b	31/.4	26/.6	22/.5
3	c	46/1.2	50/1.3	27/1.0
4	d	43/.9	49/1.5	30/1.1
5	e	53/1.1	51/0.9	29/1.2
6	f	49/1.2	55/1.4	39/.9
7	g	65/1.9	70/2.0	45/2.0
8	h	91/2.1	89/2.3	56/2.2

Table 4.2: Simulation Results

$$p_j = \frac{\left|\left(\widehat{\theta}_j - \theta_j\right)\right|}{\overline{\theta}} \times 100.$$

where $\overline{\theta}$ is the average of the θ_j's. The results shown in Table 4.2 are the mean $\overline{p}$ and standard deviation s_p defined as follows:

$$\overline{p} = \frac{1}{J}\sum_{j=1}^{J} p_j, \;\; s_p = \left[\frac{1}{J}\sum_{j=1}^{J}(p_j - \overline{p})^2\right]^{1/2}$$

From Table 4.2, we see that the Gibbs sampler estimator has the smallest percentage of error in all the simulation runs. The Rubin's estimator performs better than the QMP estimator in 3 out of 8 runs. In the first two runs, all the estimators performed relatively good. This is due to the fact that the gamma process distributions are compatible with the underlying models. However in the last run, all of them did not perform well. This run was for process distribution (h), which is an extreme two-point distribution for the θ_t's. In practice, such a distribution implies two populations, which usually can be segregated. Then the models could be applied separately for the two populations.

Chapter 5

Conclusion

Empirical and Hierarchical Bayes methods, are presently vastly underutilized in practice. This is due mainly to the fact that dealing with such methods require a lot of numerical integrations or analytic approximations in calculating the marginals of the posterior densities (often necessitating intricate attention to reparametrization and other subtleties requiring case-by-case consideration). This was the case with the Quality Measurement Plan in its first stages of implementation, where Hoadley was continuously adjusting his algorithm to cover all possible cases. For example, the first algorithm derived by Hoadley failed to consider the case of zero defects in all the rating periods. However, the excellent performance of the above methods in many real life problems, made them gain popularity and statisticians started to look for better techniques to solve these problems. The Monte Carlo-based approaches to the implementation of the Bayesian paradigm provide a flexible treatment for the calculation of marginals of the posterior density, as well as the calculation of the posterior distributions of functionals of the model parameters.

We first studied in Chapter 2 the Quality Measurement Plan as an important real life application of HB methods. This is considered as a revolution in the quality control area. Being implemented in large companies such as Bellcore justifies its practical validity. I am sure that this plan will be more and more famous in the coming years, and will be implemented in several other

companies and replaces the T-rate system. What hinders its popularity is the usage of analytical and complicated approximations in developing the QMP algorithm. So it might be difficult to convince the engineers with weak statistical backgrounds with the techniques used in developing the QMP estimator.

With the development of computing power, Monte Carlo-based approaches such as the Gibbs sampler were used to perform Bayesian computations in various important statistical problems. These methods were mainly used in the context of neural networks and expert systems, and overlooked in many conventional statistical problems. In the late 80's and in the 90's, theses approaches started to be studied and applied to some conventional statistical problems. But their applications in real life problems were quite rare. In Chapter 3, we have emphasized providing a comparative review and explication of three possible sampling approaches to the calculation of intractable marginal densities. The substitution, Gibbs, and importance-sampling algorithms are all straightforward to implement in several frequently occurring practical situations, thus avoiding numerical or analytic approximation exercises.

Next, we tried to apply these sampling approaches to the QMP model. some minor modifications are added to the QMP model in order to formulate it in HB terms ready for the application of the Monte Carlo methods. We tried to consider also computational efficiency in developing our model to be comparable to the QMP model developed by Hoadley. The simulation results at the end of this study were quite encouraging in the sense that the Gibbs estimator gave more accurate results than the previous QMP estimator. The percentage of error is reduced by at least 10 or 15 % in all the scenarios of the simulation design. The computational experience of chapter 4 also reveals that the iterative, adaptive sampling (Gibbs sampler) invariably provides better value, in terms of efficient use of generated variates, than an equivalent sample-size, noniterative, one-off approach (Rubin).

We conclude this study by highlighting some future research directions. We showed in Section 4.1.1 (Model 1), that assuming flat priors for the hyperparameters α and β results in a posterior density known just to a multiplicative

constant for α. We tried to implement the Griddy-gibbs sampler on this conditional density, but we lost a lot of computational efficiency in generating random variates from it. As explained in Chapter 3, there are more sophisticated methods that might generate faster from such densities. We can also try to look for some other plausible priors that might result in known posterior densities or at least some densities from which we can generate easily random variates.

Developing some automated convergence methods may be another important aspect of the study. This is still a hot research area in the Monte Carlo approaches such as the Gibbs sampler. In our study, we used graphical displays of the densities and functionals of the densities to demonstrate convergence of the algorithm. This may not be quite efficient if we are thinking to implement our algorithm in real life calculations. The speed of convergence may change from one input data set to another. On the one hand, if we assign the number of iterations high enough like we did in our simulation study, we may be loosing some computational effort for some data sets. On the other hand, if we assign a number of iterations barely enough to cover most of the data sets, we may be in danger of assessing the wrong distribution at the last iteration. For this, I suggest the development of an on-line, automated, interactive software to determine satisfactory convergence. The Gibbs stopper explained in Section 3.5.3 may be the best available technique for developing such a software.

Bibliography

[1] Bellcore. The Quality Measurement Plan (QMP). Technical Report TSY-000438, Bellcore, Red Bank, NJ., 1990.

[2] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society*, Ser. B(2):192–326, 1974.

[3] B. P. Carlin and A. E. Gelfand. An iterative Monte Carlo Method for Nonconjugate Bayesian Analysis. *Statistics and Computing*, 1:119–128, 1991.

[4] B. P. Carlin, A. E. Gelfand, and A. F. M. Smith. Hierarchical Bayesian Analysis of Change Point Problem. *Journal of the Royal Statistical Society*, Ser. C(41):389–405, 1991.

[5] B. P. Carlin and N. G. Poison. Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler . *Canadian Journal of Statistics*, 19:399–405, 1991.

[6] K. S. Chan. Asymptotic Behaviour of the Gibbs Sampler. *Journal of the American Statistical Association*, 88:320–326, 1993.

[7] L. Devroye. *Non-uniform Random Variate Generation.* New York: Springer-Verlag, 1986.

[8] L. Devroye and L. Gyorfi. *Non-Parametric Density Estimation: The L_1 View.* New York: John Wiley, 1985.

[9] D. K. Dey, A. E. Gelfand, and F. Peng. Overdispersed Generalized Linear Models. Technical report, University of Connecticut, Dept. of Statistics, 1994.

[10] H. F. Dodge. A Method of Rating Manufactured Product. *B.S.T.J*, 7:350–368, 1928.

[11] H. F. Dodge and M. N. Torrey. A Check Inspection and Demerit Rating Plan. *Ind. Qual. Cont.*, 13(1):1–8, 1956.

[12] B. Efron and C. Morris. Stein's Estimation Rule and its Competitors–An Empirical Bayes Approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[13] B. Efron and C. Morris. Stein's Paradox in Statistics. *Sci. Am.*, 236(5):119–127, 1977.

[14] A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. M. Smith. Illustration of Bayesian Inference in Normal Data Models using Gibbs sampling. *Journal of the American Statistical Association*, 85:972–985, 1990.

[15] A. E. Gelfand and A. F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[16] A. E. Gelfand, A. F. M. Smith, and T. M. Lee. Bayesian Analysis of Constrained Parameter and Truncated Data Problems using Gibbs sampling. *Journal of the American Statistical Association*, 87:523–532, 1992.

[17] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distirbutions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[18] W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Journal of the Royal Statistical Society*, Ser. C(41):337–348, 1992.

[19] N. Glick. Consistency Conditions for Probability Estimators and Integrals of Density Estimators. *Utilitas Mathematika*, 6:61–74, 1974.

[20] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:79–109, 1970.

[21] B. Hoadley. An Empirical Bayes Approach to Quality Assurance. In *33rd Annual Technical Conference Transactions of ASAC*, pages 257–263, 1979.

[22] B. Hoadley. The Quality Measurement Plan. *Bell System Technical Journal*, (60):215–271, 1981.

[23] B. Hoadley. QMP Theory and Algorithms. Technical Report TSY-000238, Bellcore, Red Bank, NJ., 1984.

[24] J. P. Hobert and G. Casella. Functional Compatibility, Markov Chains and Gibbs Sampling with Improper Posteriors. Technical report, University of Florida, Dept. of Statistics, 1995.

[25] J. P. Hobert and G. Casella. The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, 91:1461–1473, 1996.

[26] J. G. Ibrahim and P. W. Laud. On Bayesian Analysis of Generalized Linear Models Using Jefferey's Prior. *Journal of the American Statistical Association*, 86:981–986, 1991.

[27] K. H. Li. Imputation Using Markov Chains. *Journal of Statistical Computation and Simulation*, 30:57–79, 1988.

[28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[29] F. Nebebe and T. W. F. Stroud. Bayes and Empirical Bayes Shrinkage Estimation of Regression Coefficients: A Cross-Validation Study. *Journal of Educational Studies*, 14(4):199–213, 1988.

[30] A. E. Raftery and S. M. Lewis. How many Iterations in the Gibbs Sampler. *Bayesian Statistics*, 4:763–773, 1992.

[31] L. Rall. *Computational Solution of Non-linear Opeartor Equations.* New York: John Wiley, 1969.

[32] B. Ripley. *Stochastic Simulation.* New York: John Wiley, 1987.

[33] C. Ritter and M. A. Tanner. Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, 87:861–868, 1992.

[34] C. P. Robert. Convergence Control Methods for Markov Chain Monte Carlo Algorithms. *Statistical Science*, 10:231–253, 1995.

[35] G. O. Roberts. Convergence Diagnostics of the Gibbs Sampler. *Bayesian Statistics*, 4:775–782, 1992.

[36] G. O. Roberts and A. F. M. Smith. Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithm. *Stochastic Processes and their Applications*, 49:207–216, 1994.

[37] D. B. Rubin. Using the SIR Algorithm to Simulate Posterior Distributions. *Bayesian Statistics*, 3:395–402, 1988.

[38] D. R. Rubin. Comment on "The Calculation of Posterior Distributions by Data Augmentation," by M. A. Tanner and W.H. Wong. *Journal of the American Statistical Association*, 82:543–546, 1987.

[39] M. J. Schervish and B. P. Carlin. On the Convergence of Successive Substitution Sampling. Technical report, Carnegie Mellon University, Dept. of Statistics, 1990.

[40] W. A. Shewhart. *Economic Control of Qualityof Manufactured Product.* New York: D. Van Nostrand, 1931.

[41] W. A. Shewhart. Nature and Origin of Standards of Quality. *B.S.T.J.*, 37(1):1–22, 1958.

[42] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall, 1986.

[43] C. Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *The Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206. California: University of California Press, 1955.

[44] M. A. Tanner and W. H. Wong. The Calculation of Posterior Distributions by Data Augmentation (With Discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

[45] I. Verdinelli and L. Wasserman. Bayesian Analysis of Outlier Problems Using the Gibbs Sampler. *Statistics and Computing*, 1:105–117, 1991.

[46] C. S. Wang, J. J. Rutledge, and D. Gionola. Marginal Inferences About Variance Components in a Mixed Linear Model Using Gibbs Sampling. *Genetique, Selection, Evolution*, 25:41–62, 1993.

[47] G. C. G. Wei and M. A. Tanner. Posterior Computations with Censored Regression Data. *Journal of the American Statistical Association*, 85:829–839, 1990.

[48] A. Zaman. Cutting Corners: Efficient Random Number Generation from Unimodal Densities. Technical report, Lahor University of Management Science, Lahor, Pakistan, 1995.

[49] S. Zeger and M. R. Karim. Generalized Linear Models with Random Effects: A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86:79–86, 1991.

Vitae

Faker Zouaoui was born in 1972. He studied high school at the English Pioneer School in Ariana, Tunisia. He holds an M.S. and B.S. in Industrial Engineering from Bilkent University, Turkey. He worked as a teaching assistant in the Industrial Engineering Department of Bilkent University for two years.

His current research interests include applied probability and statistics, stochastic processes, and Bayesian statistics.